Anisotropy in Fitness Landscapes

By

Peter F. Stadler[a,b,*] and Walter Grüner[a,c]

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

* Correspondence to:
Peter F. Stadler, Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Vienna, Austria
Phone: [431] 40 480 / 678
Fax: [431] 40 28 525 151
Email: `studla@itc.univie.ac.at`

Abstract

A definition of empirical anisotropy is proposed, which allows for a quantitative measurement. This theory is applied to RNA free energy landscapes. It is shown that the biophysical **GCAU** landscapes are highly anisotropic, while the synthetic **GCXK** landscapes become isotropic for long chains. The major part of the anisotropy of the **GCAU** landscapes arises from the difference of the **G**≡**C** and the **A**=**U** stacking parameters.

1. Introduction

Sewall Wright (1932) viewed evolution as optimization process on a "fitness landscape". In the early seventies Eigen (1971) introduced the modern definition of fitness landscape as mapping from a discrete metric space, usually a sequence space, into the real numbers assigning to each sequence a particular fitness value. By now, fitness landscapes (or energy landscapes) are well-studied objects not only in theoretical biology (Fontana *et al.* 1987, 1989, Kauffman *et al.* 1987, 1988, Macken and Perelson 1989, Eigen *et al.* 1989, Weinberger 1990, 1991, Schuster 1991), but also in spin glass physics (Mézard *et al.* 1985) and in computer science (Lawler *et al.* 1987, Otten and van Ginneken 1989, Aarts and Korst 1990). Most of these investigations deal with landscapes which are instances of isotropic random fields.

For RNA landscapes, however, the situation is different. They are determined uniquely by a fairly small set of empirical parameters and therefore the question whether such a landscape is isotropic or not must be dealt with in detail.

This contribution is organized as follows: In the next three sections we derive a theory for anisotropic landscapes and show that instances from the Sherrington-Kirkpatrick spin glass are in fact isotropic according to our definitions. In section 5 these methods are applied to RNA free energy landscapes.

2. Correlated Landscapes

Two features are common to all fitness landscapes:

(1) A *cost function* or *fitness function* f assigns a value to each *configuration*, i.e., an energy to each spin configuration, a fitness value to each RNA sequence, a length to each tour in a traveling salesman problem.

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

The set $\mathcal{C}$ of all configurations can thus be viewed as a graph with each vertex corresponding to a configuration and an edge between any two nearest neighbors. Therefore, there is a natural metric distance $d(x, y)$ between any two configurations given by the shortest path in the graph $\mathcal{C}$. As an example consider the set of all sequences of given length n and define nearst neighbors in this set as sequences that differ only in a single position. The corresponding graph is known as *sequence space* (Rechenberg 1973) and the distance $d(x, y)$ is the well-known Hamming distance (Hamming 1950), i.e., the number of positions in which x and y differ.

We will need the sets $U_x(d) = \{y \in \mathcal{C} \mid d(x, y) = d\}$ consisting of all configurations with distance d from the configuration x. If its size $|U_x(d)|$ does not depend on x then $\mathcal{C}$ is said to be *distance degree regular* (Buckley and Harrary 1990). Since the configuration spaces of most combinatorial optimization problems (e.g., the TSP, graph bipartioning, graph matching), as well as the sequence spaces of biological macromolecules fulfill this (and even a more restrictive) symmetry requirement, we will assume distance degree regularity of $\mathcal{C}$ throughout this contribution. $|\mathcal{C}|$ stands for the total number of configurations. The probability to pick a pair of configurations (x, y) with distance $d(x, y) = d$ is given by

$$p(d) = \frac{|U(d)|}{|\mathcal{C}|^2} \tag{1}$$

In the following we will denote averages over the entire configuration space by $\langle\,.\,\rangle_{\mathcal{C}}$; the corresponding variance will be written as $\mathrm{var}_{\mathcal{C}}[\,.\,]$. By $\langle\,.\,\rangle_d$ we will denote averages over all pairs of vertices of the configuration space that have distance $d(x,y)=d$. For a formal definition of these averages and variances and some of their basic properties we refer to the appendix.

Correlation functions have proven to be very useful for the classification and analysis of fitness landscapes. They have been studied for various RNA landscapes (Fontana *et al.* 1991, 1993b, Bonhoeffer *et al.* 1993), for spin glass models (Weinberger and Stadler 1993), for Kauffman's Nk-model (Weinberger 1990, 1991), and for the most prominent combinatorial optimization problems, i.e., the travelling salesman problem (Stadler and Schnabl 1992), the graph bipartitioning problem (Stadler and Happel 1992), and the graph matching problem (Stadler 1992).

In accordance with previous investigations (Fontana *et al.* 1991, Bonhoeffer *et al.* 1993, Stadler and Happel 1992) we use the following:

Definition 1. The autocorrelation function of a landscape $(\mathcal{C}, f)$ is given by

$$\rho(d) = \frac{\langle f(x)f(y)\rangle_d - \langle f\rangle_{\mathcal{C}}^2}{\mathrm{var}_{\mathcal{C}}[f]} \tag{2}$$

Remark. Weinberger (1990, 1991) proposed using the autocorrelation function $r(s)$ of values sampled along a random walk $x_0, x_1, x_2, \dots$ on the configuration space $\mathcal{C}$. $r(s)$ and $\rho(d)$ are closely related via the geometric relaxation of the random walk on the configuration space. Let φ_{sd} denote the probability that the random walk reaches a distance d from the origin by moving s steps. Then $r(s)=\sum_d \varphi_{sd}\rho(d)$ (Weinberger 1990).

Theorem 1. Let $\mathcal{C}$ be distance degree regular. Then

$$\rho(d) = 1 - \frac{\left\langle \big(f(x)-f(y)\big)^2 \right\rangle_d}{2\mathrm{var}_{\mathcal{C}}[f]} \tag{3}$$

Using lemma 1 and lemma 2 in the appendix equ.(3) can be reduced to the definition of the autocorrelation function (2). ∎

As an immeadiate consequence of theorem 1 it can be show that the average correlation on $\mathcal{C}$ vanishes. (See lemma 3 in the appendix.)

Theorem 1 has been used recently to generalize the notion of an autocorrelation function to the mapping from sequence space to secondary structure. This more general treatment requires the replacement of squared differences by suitable squared distances (Fontana *et al.* 1993a).

3. Test Sets

Definition 2. In order to study "local" features of the cost function f on $\mathcal{C}$ we will use a system $\mathcal{B}$ of test sets with the following properties:

(1) $\mathcal{B}$ is a partition of the configuration space.

(2) $A \in \mathcal{B}$ is a connected subgraph of $\mathcal{C}$.

(3) Any two subgraphs $R, S \in \mathcal{B}$ are isomorphic.

(4) $1 \ll |A| \ll |\mathcal{C}|$.

These quite restrictive requirements are imposed to exclude any bias resulting from the different geometries of the test sets. Let $a = |A|$ and $n = |\mathcal{B}|$. Then we have $n \cdot a = |\mathcal{C}|$.

By $\nu_{\mathcal{B}}(d)$ we will denote the number of pairs of vertices $(x, y) \in A{\times}A$ with mutual distance d. Their distribution will be denoted as $p_{\mathcal{B}}(d) = \nu_{\mathcal{B}}(d)/a^2$.

Example 1. Consider the Boolean Hypercube B^{2n} of dimension $2n$. Let A_σ be the set of all strings (x, σ), the second half of which is the fixed string σ. Of course $\mathcal{B} = \{A_\sigma | \sigma \in B^n\}$ is a partition of B^{2n}, each element of which is a Boolean Hypercube of dimension n.

Example 2. Consider the sequence space T_4^n consisting of all sequences of length n, which can be built on a four letter alphabet. For example, consider the alphabet $\{$**G, C, A, U**$\}$ of nucleotides in RNA molecules. Construct $\mathcal{B}$ as follows. A_σ is the set of all sequences that have **G** or **C** at each position i with $\sigma_i = 1$ and an **A** or **U** on the remaining positions ($\sigma_i = 0$). The binary string σ therefore characterizes the "slice" A_σ which itself is a Boolean Hypercube with dimension n. Obviously $\mathcal{B}$ is a partition of $\mathcal{C}$ fulfilling the requirements of definition 2. We will use this choice of $\mathcal{B}$ to investigate RNA free energy landscapes in section 5.

We will denote mean and variance over a single test set A by $\langle\,.\,\rangle_A$ and $\mathrm{var}_A[\,.\,]$, respectively. Analogously we write $\langle\,.\,\rangle_d^A$ for an average over all pairs (x, y) with $x, y \in A$ and $d(x, y) = d$. Averages over all test sets contained in a partition $\mathcal{B}$ will denoted by $\langle\,.\,\rangle_{\mathcal{B}}$, and $\mathrm{var}_{\mathcal{B}}[\,.\,]$ will be used for the corresponding variance. The precise definitions of these averages can be found in the appendix.

Before we proceed to a definition of anisotropy we note that the values of the cost function on a test set A will in general be correlated, i.e., that

$$\sum_d p_{\mathcal{B}}(d)\rho(d) = \bar{\rho}_{\mathcal{B}} \neq 0 \tag{4}$$

This reflects the fact that the connected sets A are much smaller than the entire configuration space $\mathcal{C}$.

4. Isotropy and Anisotropy

A random field ϕ on a graph Γ is said to be homogeneous and isotropic if the mathematical expectation $\mathcal{E}[\phi(x)]$ of ϕ is the same for all vertices, and if the covariances $\mathcal{E}[\phi(x)\phi(y)]$ depend only on the distance $d(x, y)$. Since we are not dealing with an ensemble of landscapes but just with a single instance, we need to define a suitable analog to isotropy of random fields.

We suggest the following:

Definition 3. A value landscape is *empirically isotropic* if for any family $\mathcal{B}$ of test sets, fulfilling (1) through (4) in definition 2, holds

$$\left\langle \left(f(x) - f(y)\right)^2 \right\rangle_d^A \cong \left\langle \left(f(x) - f(y)\right)^2 \right\rangle_d . \tag{5}$$

Theorem 2. Suppose $(\mathcal{C}, f)$ is empirically isotropic. Then

$$\mathrm{var}_A[f] \cong \mathrm{var}_{\mathcal{C}}[f](1-\bar{\rho}_{\mathcal{B}}) \tag{6}$$

The proof of this theorem can be found in the appendix.

As a conseqence of theorem 2 and lemma 4 in the appendix we find for empirically isotropic landscapes:

$$\mathrm{var}_{\mathcal{B}}[\langle f\rangle_A] \cong \mathrm{var}_{\mathcal{C}}[f]\cdot\bar{\rho}_{\mathcal{B}}. \tag{7}$$

It seems natural to measure the anisotropy of a landscape by the extent to which the above relation is violated.

Definition 4. The dimensionless parameter

$$\alpha = \frac{\mathrm{var}_{\mathcal{B}}[\langle f\rangle_A]}{\mathrm{var}_{\mathcal{C}}[f]} - \bar{\rho}_{\mathcal{B}} \tag{8}$$

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

Example. Consider a Sherrington-Kirkpatrik spin glass with $2n$ spins. The autocorrelation function is given by (Weinberger and Stadler 1993)

$$\rho(d) = 1 - \frac{8nd-4d^2}{2n(2n-1)} \tag{9}$$

The partition $\mathcal{B}$ is obtained by fixing half of the spins within a given set $A\in\mathcal{B}$. The average correlation within any set $A\in\mathcal{B}$ is given by equation (15)

$$\begin{aligned}\bar{\rho}_{\mathcal{B}} &= \frac{1}{2^n}\sum_{d=0}^{n}\binom{n}{d}\left(1-\frac{8nd-4d^2}{2n(2n-1)}\right) = \\ &= \frac{n-1}{2(2n-1)}\end{aligned} \tag{10}$$

The results of numerical experiments are shown in table 1 and figure 2.

Table 1: Empirical anisotropy α measured for various instances of the Sherrington-Kirkpatrick spin glass model. It is known that the corresponding random field is homogeneous and isotropic. For each system size n we have investigated five independent instances.

n	$\bar{\alpha}$	σ
25	-0.001	0.012
50	-0.002	0.012
100	0.002	0.014
150	0.003	0.012
200	-0.001	0.007

We emphasize that our method measures anisotropy with respect to a certain partition of the configuration space. It is possible therefore that landscapes are isotropic with respect to a particular joice of the partition $\mathcal{B}$, while they are anisotropic with respect to an alternative partition $\mathcal{B}'$.

From a theoretical point of view it is tempting to define the “true anisotropy” of a landscape as the maximum value of $\alpha_{\mathcal{B}}$ from all partitions $\mathcal{B}$ fulfilling the requirements of definition 2. From a practical point of view, however, one would need prohibitively large computer resources to actually compute $\max_{\mathcal{B}} \alpha_{\mathcal{B}}$.

5. RNA Free Energy Landscapes

Fontana *et al.* (1987, 1989) proposed a model landscape based on the secondary structures of RNA. Subsequently, the free energy landscapes of RNA have been studied extensively (Fontana *et al.* 1991, 1993a,b, Bonhoeffer *et al.* 1993) for various nucleotide alphabets.

The free energy of a secondary structure is calculated by a dynamic programming algorithm based on (Zuker and Sankoff, 1984). We used the implementation by Walter Fontana and an updated parameter set provided by Danielle Konings, which is based on (Freier *et al.* 1986) and (Jaeger *et al.* 1989).

In this contribution we restrict attention to the natural four-letter alphabet **GCAU** and an artificial four letter alphabet where **A** and **U** are replaced by 2,4-diaminopyrimidine (**K**) and xanthine (**X**). **K** and **X** form a triple hydrogen bond and the energy contributions are similar to those of the **GC** pair (Piccirilli *et al.* 1990). Because of the lack of detailed measurements we use the **GC** parameter set also for **XK**. It is known (Fontana 1992b) that the **GCXK** landscape is much more rugged than the biophysical **GCAU** landscape.

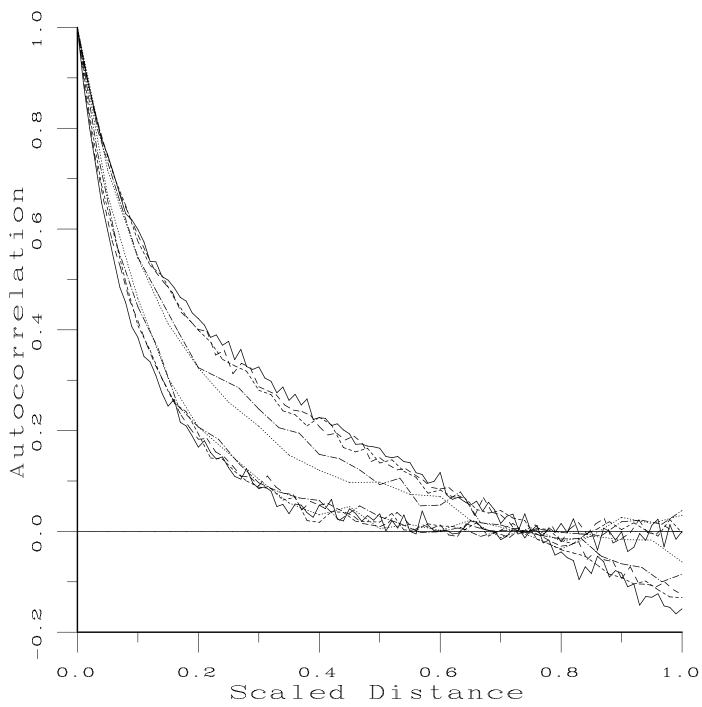

Figure 1: Empirical autocorrelation function for **GCAU** (upper set) and **GCXK** (lower set) free energy landscape as function of the scaled Hamming distance $\xi=d/n$. The line type indicates the chain length: n=20 *dotted*, n=30 *dash-dotted*, n= 50 *short dashed*, n=70 *long dashed*, n=100 *solid*. The data indicate that the autocorrelation function converge towards a limit function that is characteristic for the type of the landscape.

Table 2: Empirical coefficient of anisotropy α measured for RNA free energy landscapes. For chain length $n = 150$ and $n = 200$ the average correlation $\bar{\rho}$ of the landscape has been obtained by extrapolation.

n	**GCAU**	**GCXK**
20	0.147	0.071
30	0.158	0.048
50	0.125	0.032
70	0.178	0.019
100	0.176	0.022
150	0.199	0.010
200	0.203	0.012

Numerical experiments (figure 2, table 2) clearly reveal that the **GCAU** is highly anisotropic, while the **GCXK** landscape becomes isotropic in the limit of large chain length n.

In order to determine the major source of the empirical anisotropy in RNA free energy landscapes we consider the following simplification: Suppose the average free energy of a slice $A \in \mathcal{B}$ is only a function of the **AU**- or **XK**-content η. Assuming that slices with equal values of η are statistically indistinguishable (this is the crucial approximation!), we may estimate the coefficient of anisotropy by

$$\alpha \approx \frac{1}{\mathrm{var}_{\mathcal{C}}[f]} \cdot \left[\int \bar{F}(\eta)^2 p(\eta) d\eta - \left(\int F(\eta) p(\eta) d\eta \right)^2 \right] \tag{11}$$

After some algebra (which can be found in the appendix) we find

$$\alpha = \frac{a^2}{4n \mathrm{var}_{\mathcal{C}}[f]} + \frac{b^2 + 2ac}{32n^2 \mathrm{var}_{\mathcal{C}}[f]} + \dots \tag{12}$$

where $a = F'(\frac{1}{2})$, $b = F''(\frac{1}{2})$, and $c = F'''(\frac{1}{2})$, respectively. 208z The variances of the free energies $\mathrm{var}_{\mathcal{C}}[F]$ have been calculated accurately by Fontana *et al.* (1993) for various alphabets and chain lengths. Estimates for the parameters a and b have been extracted from the η-dependent data in fig.3 (Grüner 1992). For chain length $n = 50$ and $n = 70$ we have

$$\begin{array}{rclcrcl} \sigma^2_{50} & = & 9.6\,(\mathrm{kcal/mol})^2 & \qquad & a_{50} & = & 22.6 \text{ kcal/mol} \\ \sigma^2_{70} & = & 14.1\,(\mathrm{kcal/mol})^2 & & a_{70} & = & 34.6 \text{ kcal/mol} \end{array}$$

from the **GCAU** landscape; for the **GCXK** landscape we obtain

$$\begin{array}{rclcrcl} \sigma^2_{50} & = & 12.5\,(\mathrm{kcal/mol})^2 & \qquad & b_{50} & = & -107.2 \text{ kcal/mol} \\ \sigma^2_{70} & = & 16.8\,(\mathrm{kcal/mol})^2 & & b_{70} & = & -151.8 \text{ kcal/mol} \end{array}$$

The above data refer to fairly short chains; therefore, they cannot be expected to yield very accurate estimates for the asymptotic anisotropy parameter. The resulting values for the coefficient of anisotropy are tabulated in table 3.

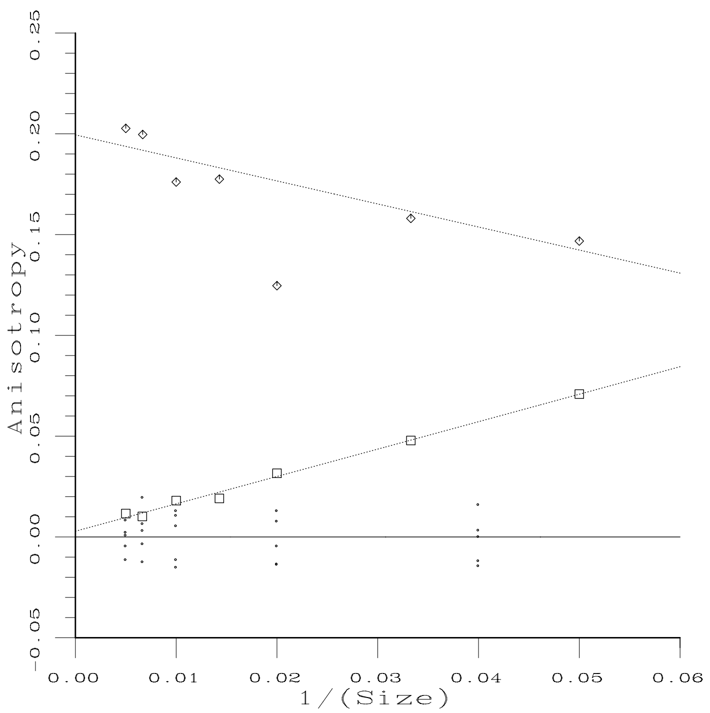

Figure 2: Emprical coefficient of anisotropy as a function of chain length. ◇ **GCAU**, □ **GCXK**, · Sherrington Kirkpatrick spin glass.

Fontana *et al.* (1991, 1993b) showed that the mean free energies as well as the variances can very well be approximated by a linear function:

$$\begin{aligned} a = \left.\frac{dF}{d\eta}\right|_{\frac{1}{2}} &\sim a_0 + a_1 n \\ b = \left.\frac{d^2F}{d\eta^2}\right|_{\frac{1}{2}} &\sim b_0 + b_1 n \\ c = \left.\frac{d^3F}{d\eta^2}\right|_{\frac{1}{2}} &\sim c_0 + c_1 n \\ \mathrm{var}_{\mathcal{C}}[f] &\sim v_0 + v_1 n \end{aligned} \tag{13}$$

Substituting this into equation (12) finally yields

$$\alpha = \frac{1}{4}\frac{a_1^2}{v_1} + \left[\frac{1}{2}\frac{a_0 a_1}{v_1} + \frac{1}{16}\frac{a_1 c_1}{v_1} + \frac{1}{32}\frac{b_1^2}{v_1} - \frac{1}{4}\frac{a_1^2 v_0}{v_1^2}\right]\frac{1}{n} + \mathcal{O}(\frac{1}{n^2}) \tag{14}$$

We expect to find two types of landscapes depending on whether $a_1 = 0$ or not. For models that are symmetric with respect to $\eta \to \eta - \frac{1}{2}$, i.e., that have $a_1 = 0$, we find

$$\alpha \approx \frac{1}{n} \cdot \frac{b_1^2}{32 v_1} \tag{15}$$

while for asymmetric landscapes we obtain

$$\alpha \approx \frac{a_1^2}{4v_1} + \mathcal{O}(\frac{1}{n}) \tag{16}$$

Obviously, **GCXK** belongs to the first type, while **GCAU** belongs to the second type.

Unfortunately, at present no dataset is available that would allow us to obtain the constants a_0, a_1 etc. with sufficient accuracy. Since the production of such data is extremely time consuming we stick to a simplified model.

Table 3: Prediction of anisotropy for RNA landscape from simple analytical models.

	equ.(28)	equ.(35)	exp
GCAU	0.49	0.28	0.20
GCXK	$1.29/n$	$0.60/n$	$1.36/n$

Let P_{GC} be the probability that at two randomly picked positions in a random sequence with **AU**- or **GC**-content η there are a **C** and a **G**. It is easy to check that

$$P_{GC} = \frac{1}{2}(1-\eta)^2, \qquad P_{AU} = \frac{1}{2}\eta^2, \qquad P_{GU} = \frac{1}{2}\eta(1-\eta) \tag{17}$$

Assuming that the loop strain energies do not strongly depend on the alphabet, the average free energy of the structures in a slice with an **AU**- or **XK**- content η may be approximated by

$$\begin{aligned} \bar{F}_{\mathbf{GCAU}}(\eta) &= 2F[AU]P_{AU} + 2F[GC]P_{GC} + 2F[GU]P_{GU} = \\ &= \eta^2\left(F[AU] + F[GC] - F[GU]\right) - \\ &\qquad - \eta\left(2F[GC] - F[GU]\right) + F[GC] \end{aligned} \tag{18}$$

$$\begin{aligned} \bar{F}_{\mathbf{GCXK}}(\eta) &= 2F[GC]P_{GC} + 2F[XK]P_{XK} = \\ &= F[GC](2\eta^2 - 2\eta + 1) \end{aligned}$$

yielding the estimates

$$\begin{aligned} &a_{\mathbf{GCXK}} = 0 \qquad\qquad b_{\mathbf{GCXK}} = 4F[GC] \\ &a_{\mathbf{GCAU}} = F[AU] - F[GC] \end{aligned} \tag{19}$$

Numerical simulations (Fontana *et al.* 1992b) yield

$$\begin{aligned} F[GC] &= 10.12 - 0.781n \text{ kcal/mol} \\ F[AU] &= 2.61 - 0.109n \text{ kcal/mol} \\ F[GU] &= 0.27 - 0.010n \text{ kcal/mol} \\ \mathrm{var}[GCAU] &= -1.80 + 0.226n \text{ (kcal/mol)}^2 \\ \mathrm{var}[GCXK] &= -0.16 + 0.236n \text{ (kcal/mol)}^2 \end{aligned}$$

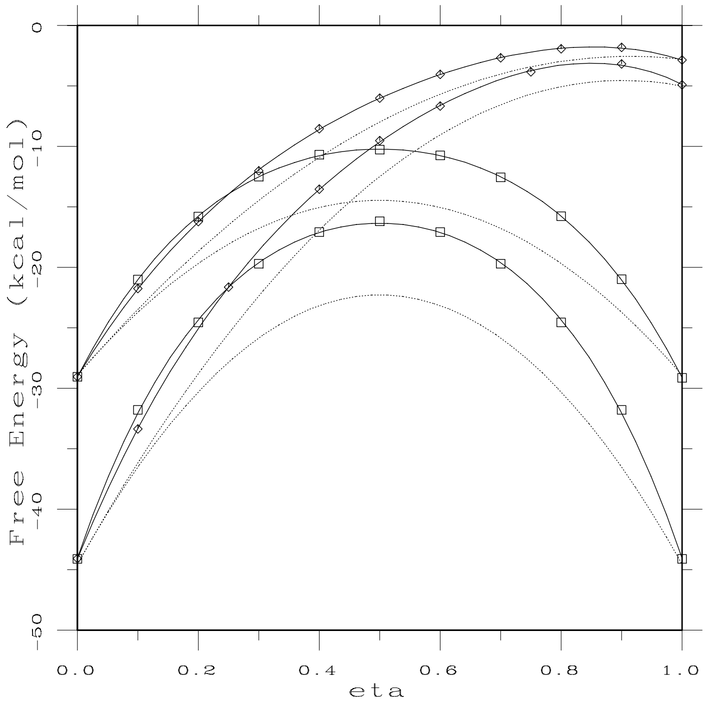

Figure 3: Average free energies as function of the **GC** content. $\diamond$ **GCAU**, $\square$ **GCXK**. Solid lines are quartic fits that have been used to calculated the parameters a and b. Dotted lines refer to the model equation (34). Upper curves n=50, lower curves n=70.

The resulting estimates for the coefficients of anisotropy are compiled in table 3.

6. Conclusions

We have shown that a measure for the anisotropy of fitness landscapes can be defined in a canonical way. It is based entirely on second moments. Instances from isotropic random fields such as the Sherrington-Kirkpatrick spin glass are, as expected, isotropic. The theory presented in this contribution is capable only of detecting that a landscape is anisotropic. No detailed information on the topology of the anisotropies can be gained.

Application of our theory to RNA free energy landscapes shows a striking difference between the biophysical **GCAU**-alphabet and an artificial **GCXK** landscape with two base pairs of equal strength. While the **GCXK** landscape becomes isotropic for large chain lengths, the biophysical landscape remains highly anisotropic. The main reason for this behaviour is the difference of the energy parameters for **GC** and **AU** pairs, respectively. Simplified models taking into account only that the average free energy depends on the **GC** content allow for semi-quantitative estimates of the coefficients of anisotropy. Both landscapes again exhibit the qualitative difference between them.

The spin glass models and combinatorial optimization problems commonly used for studies on the dynamics of evolution processes are isotropic. It is not straightforward therefore to apply results obtained from such model landscapes to the presumably higly anisotropic "natural" fitness landscape. While for isotropic landscapes the autocorrelation function $\rho(d)$ seems to be the crucial quantity for performance of optimization algorithms (and for the velocity of evolution), it is likely that on highly anisotropic landscapes the particular geometric details of the anisotropies are most important. It has been pointed out, for instance (Eigen *et al.* 1989), that ridge-like structures may speed up the evolution of a quasispecies by orders of magnitudes, while net-like structures can stabilize a broadly distributed quasispecies (Tarazona 1992). The influence of anisotropies on the performance of various types of optimization algorithms deserves detailed research in future.

Acknowledgements

Useful discussions with Peter Schuster and Walter Fontana are greatfully acknowledged. The spin glass data were calculated on the IBM-ES/9021-720 mainframe of the Computer Center of the University of Vienna, who generously supplied the computer time in connection with the IBM-EASI program. Thanks also to Pam Amick, who proofread this manuscript.

References

[1] Aarts E.H.L. and J. Korst. Simulated Annealing and Boltzman Machines. J.Wiley & Sons, New York, 1990.

[2] Bonhoeffer S., J.S. McCaskill, P.F. Stadler, and P. Schuster. RNA Multistructure Landscapes. A Study Based on Temperature Dependent Partition Functions. *Eur. Biophys. J.* in press (1993).

[3] Buckley F. and F. Harary. Distance in Graphs. Addison Wesley, Redwood City, Cal. 1990.

[4] Eigen M. Selforganization of Matter and the Evolution of Macromolecules. *Naturwissenschaften* **10**, 465-523 (1971).

[5] Eigen M., J.S. McCaskill, and P. Schuster. The Molecular Quasispecies. *Adv. Chem. Phys.* **75**, 149-263 (1989).

[6] Fontana W. and P. Schuster. A Computer Model of Evolutionary Optimization. *Biophys. Chem.* **26**, 123-147 (1987).

[7] Fontana W., W. Schnabl, and P. Schuster. Physical Aspects of Evolutionary Optimization and Adaptation. *Phys. Rev. A* **40**, 3301-3321 (1989).

[8] Fontana W., T. Griesmacher, W. Schnabl, P.F. Stadler, and P. Schuster. Statistics of Landscapes Based on Free Energies Replication and Degradation Rate Constants of RNA Secondary Structures. *Mh. Chem.* **122**, 795-819 (1991).

[9] Fontana W., D.A.M. Konings, P.F. Stadler, and P. Schuster. Statistics of RNA Secondary Structures. *SFI preprint* 92-02-007, *Biopolymers* in press (1993a).

[10] Fontana W., P.F. Stadler, E. Bauer, T. Griesmacher, I.L. Hofacker, M. Tacker, P. Tarazona, E.D. Weinberger, and P. Schuster. RNA Folding and Combinatory Maps. *Phys. Rev. E* in press (1993b).

[11] Freier S.M., 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. *Biochemistry* **83**, 9373-9377 (1986).

[12] Grüner W. Complex Combinatory Maps of RNA Secondary Structures: A Systematic Study on Base Composition Effects. *Diploma Thesis*, University of Vienna 1992.

[13] Hamming R.W. Error Detecting and Error Correcting Codes. *Bell Syst. Tech. J.* **29**, 147-160 (1950).

[14] Jaeger J.A., D.H. Turner, and M. Zuker. Improved Predictions of Secondary Structures for RNA. *Biochemistry* **86**, 7706-7710 (1989).

[15] Kauffman S.A. and S. Levin. Towards a General Theory of Adaptive Walks on Rugged Landscapes, *J. theor. Biol.* **128**, 11-45 (1987).

[16] Kauffman S.A., E.D. Weinberger, and A.S. Perelson. Maturation of the Immune Response Via Adaptive Walks On Affinity Landscapes, *Theoretical Immunology, Part I Santa Fe Institute Studies in the Sciences of Complexity*, A.S. Perelson (ed.), Addison-Wesley, Reading, Ma. 1988.

[17] Lawler E., A. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys. The Travelling Salesman Problem. Wiley, New York 1985.

[18] Macken C.A. and A.S. Perelson. Protein Evolution on Rugged Landscapes, *Proceedings of the National Academy of Sciences* **86**, 6191-6195 (1989).

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

[20] Otten R.H.J.M and L.P.P.P. van Ginneken. The Annealing Algorithm. Kluwer Academic Publ., Boston 1989.

[21] Piccirilli J.A., T. Krauch, S.E. Moroney, and S.A. Brenner. Enzymatic Incorporation of a New Base Pair into DNA and RNA Extends the Genetic Alphabet. *Nature* **343**, 33 (1990).

[22] Rechenberg I. Evolutionsstrategie. Frommann-Holzboog, Stuttgart, Germany, 1973.

[23] Schuster P. Complex Optimization in an Artificial RNA World. In: D. Farmer, C. Langton S. Rasmussen and C. Taylor: *Artificial Life II*, SFI Studies in the Science of Complexity XII Addison-Wesley, Reading, Mass. 1991.

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

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

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

[27] Tarazona P. Error Threshold for Molecular Quasispecies as Phase Transition. *Phys. Rev. A[15]* **45**, 6038-6050 (1992).

[28] Turner D.H., N. Sugimoto, and S. Freier. RNA Structure Prediction. *Ann. Rev. Biophys. Biophys. Chem.* **17**, 167-192 (1988).

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

[30] Weinberger, E.D. Local Properties the N-k Model, a Tuneably Rugged Energy Landscape. *Phys. Rev. A* **44**, 6399-6413 (1991).

[31] Weinberger E.D. and P.F. Stadler. Why Some Fitness Landscapes are Fractal. *J. Theor. Biol.* in press (1993).

[32] Wright S. The Role of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution. In: *Proceedings of the 6th International Congress on Genetics* Vol.1, pp. 356-366. (1932).

[33] Zuker M. and D. Sankoff. RNA Secondary Structures and Their Prediction. *Bull. Math. Biol.* **46**, 591-621 (1984).

Appendix

This appendix contains a number of mathematical technicalities that are necessary to proof the main statements of this paper.

Notation. The mean and variance of the cost function f over the entire configuration space $\mathcal{C}$ are defined as

$$\begin{aligned}\langle f\rangle_{\mathcal{C}} &= \frac{1}{|\mathcal{C}|}\sum_{x\in\mathcal{C}} f(x)\\ \mathrm{var}_{\mathcal{C}}[f] &= \langle f^2\rangle_{\mathcal{C}} - \langle f\rangle_{\mathcal{C}}^2.\end{aligned} \tag{A.1}$$

$|\mathcal{C}|$ denotes the total number of configurations. For any function $h(x,y)$ we define

$$\langle h(x,y)\rangle_{\mathcal{C}} = \frac{1}{|\mathcal{C}|^2}\sum_{x\in\mathcal{C}}\sum_{y\in\mathcal{C}} h(x,y) \tag{A.2}$$

For an arbitrary function $h:\mathcal{C}\times\mathcal{C}\to\mathbb{R}$ let

$$\langle h(x,y)\rangle_d = \frac{1}{N\cdot|U(d)|}\sum_{x\in\mathcal{C}}\sum_{y\in U_x(d)} h(x,y) \tag{A.3}$$

The basic properties of these averages and variances are summarized in the following lemmata:

Lemma 1.

$$\mathrm{var}_{\mathcal{C}}[f] = \frac{1}{2}\Big\langle\big(f(x)-f(y)\big)^2\Big\rangle_{\mathcal{C}} \tag{A.4}$$

Proof.

$$\Big\langle\big(f(x)-f(y)\big)^2\Big\rangle = 2\frac{1}{|\mathcal{C}|}\sum_{x\in\mathcal{C}} f^2(x) - 2\frac{1}{|C|}\sum_{x\in\mathcal{C}} f(x)\cdot\frac{1}{|C|}\sum_{y\in\mathcal{C}} f(y) = 2\mathrm{var}_{\mathcal{C}}[f]$$

■

Lemma 2. Suppose that $h(x,y)=\eta(x)$ and $\mathcal{C}$ is distance degree regular. Then

$$\langle\eta(x)\rangle_d = \langle\eta(y)\rangle_d = \langle\eta\rangle_{\mathcal{C}} \tag{A.5}$$

Proof. Since for each $y\in\mathcal{C}$ there are exactly $|U(d)|$ vertices x such that $y\in U_x(d)$. As a consequence we have for arbitrary $h(x,y)$

$$\sum_{x\in C}\sum_{y\in U_x(d)} h(x,y) = \sum_{y\in C}\sum_{x\in U_y(d)} h(x,y)$$

Observing that $\sum_{y\in U_x(d)} 1 = |U(d)|$ completes the proof. ■

Lemma 3.

$$\sum_d p(d)\rho(d) = 0 \tag{A.6}$$

Proof. $\sum_d p(d)\big\langle\big(f(x)-f(y)\big)^2\big\rangle_d = \big\langle\big(f(x)-f(y)\big)^2\big\rangle_{\mathcal{C}} = 2\mathrm{var}_{\mathcal{C}}[f]$ ∎

Notation. For a function $f:\mathcal{C}\to\mathbb{R}$ we define mean and variance for the restriction to $A\in\mathcal{B}$ by

$$\begin{aligned}\langle f\rangle_A &= \frac{1}{a}\sum_{x\in A} f(x)\\ \mathrm{var}_A[f] &= \langle f^2\rangle_A - \langle f\rangle_A^2\end{aligned} \tag{A.7}$$

For any function $q:\mathcal{B}\to\mathbb{R}$ we define

$$\begin{aligned}\langle q\rangle_{\mathcal{B}} &= \frac{1}{n}\sum_{A\in\mathcal{B}} q(A)\\ \mathrm{var}_{\mathcal{B}}[q] &= \langle q^2\rangle_{\mathcal{B}} - \langle q\rangle_{\mathcal{B}}^2\end{aligned} \tag{A.8}$$

For a function $h:A\times A\to\mathbb{R}$ we define

$$\begin{aligned}\langle h(x,y)\rangle_A &= \frac{1}{a^2}\sum_{x,y\in A} h(x,y)\\ \langle h(x,y)\rangle_d^A &= \frac{1}{a\cdot\nu(d)}\sum_{x\in A}\ \sum_{y\in[A\cap U_x(d)]} h(x,y)\end{aligned} \tag{A.9}$$

Lemma 4. The quantities above fulfil

$$\begin{aligned}\langle\,\langle f\rangle_A\,\rangle_{\mathcal{B}} &= \langle f\rangle_{\mathcal{C}}\\ \mathrm{var}_{\mathcal{C}}[f] &= \langle \mathrm{var}_A[f]\rangle_{\mathcal{B}} + \mathrm{var}_{\mathcal{B}}[\langle f\rangle_A]\end{aligned} \tag{A.10}$$

Proof. The first line is an immediate consequence of the definition. For the variance we have

$$\begin{aligned}\mathrm{var}_{\mathcal{C}}[f] &= \langle\,\langle f^2\rangle_A\,\rangle_{\mathcal{B}} - \langle\,\langle f\rangle_A\,\rangle_{\mathcal{B}}^2 =\\ &= \langle\,\langle f^2\rangle_A - \langle f\rangle_A^2\rangle_{\mathcal{B}} + \langle\,\langle f\rangle_A^2\,\rangle_{\mathcal{B}} - \langle\,\langle f\rangle_A\,\rangle_{\mathcal{B}}^2\end{aligned}$$

The definition of the variance with respect to A and $\mathcal{B}$ completes the proof.

∎

We also need a more convienient representation of the variance of f on a test set A.

Lemma 5.

$$\mathrm{var}_A[f] = \frac{1}{2}\sum_d p_{\mathcal{B}}(d)\big\langle\big(f(x)-f(y)\big)^2\big\rangle_d^A \tag{A.11}$$

Proof.

$$\begin{aligned}
\langle(f(x)-f(y))^2\rangle_A &= \frac{1}{a^2}\sum_{x\in A}\sum_{y\in A}(f(x)-f(y))^2 = \\
&= \sum_d \frac{\nu(d)}{a^2}\cdot\frac{1}{\nu(d)}\sum_{x\in A}\sum_{y\in[A\cap U_x(d)]}(f(x)-f(y))^2 = \\
&= \sum_d p_{\mathcal{B}}(d)\langle(f(x)-f(y))^2\rangle_d^A
\end{aligned}$$

■

The results stated above are used to prove the main theorem of this paper:

Theorem 2. Suppose $(\mathcal{C}, f)$ is empirically isotropic. Then

$$\mathrm{var}_A[f] \cong \mathrm{var}_{\mathcal{C}}[f](1-\bar{\rho}_{\mathcal{B}}) \tag{A.12}$$

Proof.

$$\begin{aligned}
\mathrm{var}_A[f] &= \frac{1}{2}\sum_d p_{\mathcal{B}}(d)\big\langle\big(f(x)-f(y)\big)^2\big\rangle_d^A \\
&\cong \frac{1}{2}\sum_d p_{\mathcal{B}}(d)\big\langle\big(f(x)-f(y)\big)^2\big\rangle_d \\
&= \mathrm{var}_{\mathcal{C}}[f](1-\bar{\rho}_{\mathcal{B}})
\end{aligned}$$

■

The remaining part of the appendix consist of the derivation of equ.(12): In order to evaluate $\bar{F}(\eta)$ in equ.(11) we average over a sample of slices with "reference points" which are randomly distributed in $\mathcal{C}$. Thus we expect $\bar{\rho} = 0$. The distribution of the **AU**- or **XK**-content η in the set of all sequences of length n is given by the binomial distribution, which can be approximated by a Gaussian distribution.

$$p(\eta) = \frac{1}{2^n}\binom{n}{n\eta} \approx \sqrt{\frac{2}{n\pi}}\exp\left[-2n(\eta-\tfrac{1}{2})^2\right] \tag{A.12}$$

Using the Taylor expansion

$$F(\eta) = F_0 + (\eta-\tfrac{1}{2})a + \frac{1}{2}(\eta-\tfrac{1}{2})^2 b + \frac{1}{6}(\eta-\tfrac{1}{2})^3 c + \dots \tag{A.13}$$

and the notation $\mu_k = \int(\eta-\frac{1}{2})^k d\eta$ for the k-th central moment we may evaluate the integrals in equation (11) explicitly:

$$\begin{aligned} \int F(\eta)p(\eta)d\eta &= F_0 + \frac{1}{2}\mu_2 b + \frac{1}{24}\mu_4 d + \dots \\ \int F^2(\eta)p(\eta)d\eta &= F_0 + \mu_2[b\bar{F} + a^2] + \mu_4[\frac{1}{12}d\bar{F} + \frac{1}{3}ac + \frac{1}{4}b^2] + \dots \end{aligned} \tag{A.14}$$

Hence we obtain for the bracketed expression in equ.(11), which may be interpreted as variance with respect to η:

$$\mathrm{var}_\eta[f] = \mu_2 a^2 + \frac{1}{4}[\mu_4 - \mu_2^2]b^2 + \frac{1}{3}\mu_4 ac + \mathcal{O}(\mu_6) \tag{A.14}$$

Using the Gaussian approximation of the binomial distribution we have $\mu_k = 0$ for odd k and

$$\mu_2 = \sigma^2 = \frac{1}{4n} \qquad \mu_4 = 3\sigma^4 = \frac{3}{16n^2} \qquad \mu_{2k} = \mathcal{O}(\frac{1}{n^k}) \tag{A.15}$$

Substituting this into equ.(11) yields immediately the desired formula (12).

Figure Captions

Figure 1. Empirical autocorrelation function for **GCAU** (upper set) and **GCXK** (lower set) free energy landscape as function of the scaled Hamming distance $\xi = d/n$. The line type indicates the chain length: $n = 20$ *dotted*, $n = 30$ *dash-dotted*, $n = 50$ *short dashed*, $n = 70$ *long dashed*, $n = 100$ *solid*. The data indicate that the autocorrelation function converge towards a limit function that is characteristic for the type of the landscape.

Figure 2. Emprical coefficient of anisotropy as a function of chain length. $\diamondsuit$ **GCAU**, $\square$ **GCXK**, $\cdot$ Sherrington Kirkpatrick spin glass.

Figure 3. Average free energies as function of the **GC** content. $\diamondsuit$ **GCAU**, $\square$ **GCXK**. Solid lines are quartic fits that have been used to calculated the parameters a and b. Dotted lines refer to the model equation (34). Upper curves $n = 50$, lower curves $n = 70$.