Why *Some* Fitness Landscapes are Fractal

By

EDWARD D. WEINBERGER[a,†] AND PETER F. STADLER[a,b,c,*]

[a]Max Planck Institut für Biophysikalische Chemie
080 Biochemische Kinetik
Am Faßberg, D-3400 Göttingen, Germany

[b]Institut für Theoretische Chemie
Universität Wien
Währingerstraße 17, A-1090 Vienna, Austria

[c]Santa Fe Institute
1660 Old Pecos Trail, Suite A, Santa Fe NM 87501

[†]Mailing Address: 370 Central Park West, Apt. 409 New York, NY 10025, USA

*Correspondence to: Peter F. Stadler
Institut für Theoretische Chemie, Univ. Wien
Währingerstraße 17, A-1090 Vienna, Austria
Phone: **43-1-40480/678, Fax **43-1-4028525
E-mail: studla@tiger.itc.univie.ac.at

Abstract

Many biological and biochemical measurements, e.g. the "fitness" of a particular genome, or the binding affinity to a particular substrate, can be treated as a "fitness landscape", an assignment of numerical values to points in sequence space (or some other configuration space). As an alternative to the enormous amount of data required to completely describe such a landscape, we propose a statistical characterization, based on the properties of a random walk through the landscape, and, more specifically, its autocorrelation function. Under assumptions roughly satisfied by two classes of simple model landscapes (the N-k model and the p-spin model) and by the landscape of estimated free energies of RNA secondary structures, this autocorrelation function, along with the mean and variance of individual points and the size of the landscape, completely characterize it. Having noted that these and other landscapes of estimated replication and degradation rates all have a well defined correlation length, we propose a classification of landscapes depending on how the correlation length scales with the diameter of the landscape. The landscapes of some of the kinetic parameters of RNA molecules scale similarly to the model landscapes introduced into evolutionary studies from other fields, such as quadratic spin glasses and the traveling salesman problem, but the correlation length of RNA landscapes are considerably smaller. Nevertheless, both the model and some of the RNA landscapes satisfy a test of self-similarity proposed by Sorkin (1989).

Towards a statistical characterization of fitness landscapes

The notion of a "value" or "fitness landscape" is one of the most powerful concepts in evolutionary theory. Implicit in this idea — which, by now, has been discussed by many authors, beginning with Wright (1932), and including Maynard-Smith (1970), Eigen (1971, 1985), Eigen & Schuster (1978), Schuster & Swetina (1982), Schuster (1986, 1987), Fontana *et al.* (1987, 1988), Kauffman & Levin (1987), Eigen *et al.* (1989), and Kauffman & Weinberger (1989), Bak *et al.* (1992), Flyvbjerg & Lautrup (1992) —

Figure 1: Sequence Spaces for binary sequences of length N=3 and N=5. All pairs of sequences with Hamming distance 1 are connected by a line. The graph defined in this way is a hypercube of dimensions N. Boolean Hypercubes are the configuration spaces for the N-k model, the p-spin model and the **GC**-only RNA model.

is a collection of genotypes arranged in an abstract "sequence space", with each genotype "next to" those other genotypes differing from the original by a single mutation. This space and the numerical fitness values assigned to each genotype via some measurable phenotypic characteristic constitutes the fitness landscape. One of the most natural applications of this metaphor is to describe the "fitness" of individual molecules in performing a specific function, such as (a protein) binding to a particular ligand, or RNA replication *in vitro*. Other examples, the ones of greatest interest in this paper, are the free energies of RNA secondary structure folding and the rates of RNA replication and degradation.

The geometrical structure underlying each of these examples is a discrete regular lattice whose dimension, ν, is the same as the length of the sequence. Each point in the most realistic model of an RNA landscape would have 3ν neighbors, one for each of the 3 bases that could be substituted at each site. Such lattices are, of course, extremely difficult to visualize, but, fortunately, the essential features of these landscapes are reflected in the much simpler geometry of the Boolean hypercube shown in Figure 1. In fact, precisely this geometry — the lattice formed by the vertices of a ν dimensional cube — is obtained when an RNA sequence is assumed to consist either of unspecified purines and pyrimidines, rather than specific bases, or if pure **GC**-sequences are considered. In either scenario, the molecule is essentially being modeled by a string of "bits", each of which can only assume two discrete values, which we might as well label "0" and "1".

It has long been known that evolutionary dynamics on a given landscape depends crucially on its statistical structure. For example, Eigen *et al.* (1989) argue that the standard Darwinian assumption of random mutation must be re-examined for landscapes in which fitness values are not distributed randomly in sequence space. Instead, fitness landscapes with the distinct "mountainous" and flat regions familiar from terrestrial landscapes will sometimes have "ridges" of high fitness connecting pairs of local optima. Even if *any single individual* on such a ridge has the same probability of mutating as a single individual in the lowlands, their difference in fitness and thus in frequency guarantees that more mutant copies of individuals on the ridge will appear. The fitness difference between individuals on and off the ridge will amplify this disparity, and the population will tend to flow along the ridge. The resulting stationary population distribution can therefore contain "fingers" that protrude long distances in sequence space, possibly as far as another local optimum. If the known replication and mutation rates for viruses are used in numerical estimates of populations on and off the ridges, the individuals populating the "fingers" are many orders of magnitude more numerous.

As a first step towards describing the geometry of RNA fitness landscapes, Eigen *et al.* (1989) proposed that they were "self-similar". The purpose of the present paper is to understand this claim in the context of a general mathematical theory, and thus show the precise sense in which the claim is true for model RNA fitness landscapes. Our approach involves the study of correlations between the fitnesses of various points on the landscape. The need for a statistical description of what is, in principle, a deterministic phenomenon is explained in the following paragraphs.

The most obvious method of specifying a fitness distribution, which would involve simple enumeration of the fitnesses of each genotype, is inadequate because the number of genotypes — and therefore the amount of data required for the specification — grows exponentially with the size of the landscape (Even for the simple ν dimensional binary landscapes discussed above, there are 2^ν vertices on the hypercube.) Simple enumeration also fails to capture the interrelationships between neighboring fitness values. It is therefore preferable to characterize landscapes by easily measurable *statistical* properties. The conceptual foundation behind this approach is the view that the landscape is just one exemplar chosen from an ensemble of possible landscapes, an approach that generalizes the notion of a random variable to a *random landscape.* We adopt this approach for much the same reasons that the ensemble concept is used in statistical thermodynamics: the measurement of every free energy on the landscape — like the deterministic description of individual atoms in a gas — would require an unperformable number of measurements. Furthermore, such detailed information does us no good even if we had it because we are really interested in global, rather than local properties.

The statistical approach has a long history in condensed matter physics, in which it has been used to model the properties of alloys with unusual magnetic properties known as spin glasses. One such model is the Sherrington-Kirkpatrick (SK) (Sherrington & Kirkpatrick 1975), which specifies the Hamiltonian, $\mathcal{H}$, i.e. the total energy of the spin vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$. Here, it is customary to define the individual spins $\sigma_i \in \{-1, +1\}$ (rather than 0 and 1). The explicit form of the model is

$$\mathcal{H}(\sigma) = \sum_{ij} J_{ij}\sigma_i\sigma_j,$$

where the coefficients J_{ij} are independent random samples from some probability distribution. In the simplest case, the mean of the J's is zero and their variance is $\mathcal{O}(N^{-1})$, thus guaranteeing that typical values of $\mathcal{H}$ are $\mathcal{O}(1)$.

Anderson (1983) proposed that RNA landscapes are "in the same universality class" as spin-glass models, and, in particular, the SK-model. Their proposal was based on the observation that the SK model has an important property that might be expected of biological fitness landscapes — that of conflicting constraints on fitness (i.e. "frustration"), leading to many local fitness maxima. This observation also leads to the interchangeability of the biological term "fitness landscape" and the physical term "energy landscape" in this paper.

However, the crucial assumption has been made in the formulation of spin glass models, in general, and the SK model in particular; namely, that the energy "landscape" is determined primarily by *pairwise* interactions between spins. This need not — and, as we will see below, is not — true for RNA fitness landscapes. We must therefore consider a broader class of models.

There are both theoretical and empirical reasons to assume that fitnesses on the landscape have a joint Gaussian distribution. From the theoretical point of view, the fitnesses are determined by adding a large number of relatively independent factors of approximately equal strength. Under surprisingly general assumptions — even more general than the usual hypothesis of statistical independence — the Central Limit Theorem guarantees that such sums are approximately normal. (See, for example, Breiman 1986) The free energies and replication rates of randomly chosen sequences in plausible RNA landscapes appear to be normally distributed for sufficiently large chain length ν, a result to be discussed in more detail in Fontana *et al.* (1993).

At first glance, the Gaussian assumption complicates matters, because estimates of $\mathcal{O}(N^2)$ covariances, as well as estimates of the N means must be made. An additional assumption is required. The simplest possible assumption is that all of the fitnesses are *independent* Gaussian random variables, and that no additional information about a fitness value can be gleaned from a knowledge of neighboring fitnesses. This landscape, known to the physics community as the "random energy model" (Derrida 1981), has received considerable attention in recent years (Kauffman & Levin 1987, Kauffman *et al.* 1988, Kauffman & Weinberger 1989, Weinberger 1990, Bak *et al.* 1992, Flyvbjerg & Lautrup 1992). However, such landscapes are unrealistically "rugged"; in most practical situations, there is almost certainly *some* correlation between neighboring points.

The next step up in complexity is the so-called "AR(1) landscape", in which all correlations are completely determined by the (equal) correlations between neighboring points (Weinberger 1990, 1991a,b,c, 1992), that is, according to the formula

$$F(\mathbf{p}') = \rho F(\mathbf{p}) + \Delta F,$$

where ρ is the correlation between the neighboring points $\mathbf{p}$ and $\mathbf{p}'$, $F(\mathbf{p})$ and $F(\mathbf{p}')$ are their respective fitnesses, and the random variable ΔF is chosen independently from the same Gaussian distribution for each pair of points. This condition guarantees that the sequence of fitnesses obtained by a random walk in any direction will have the same statistical properties; hence, such a landscape is "statistically isotropic".

Because all of the correlation information about a given fitness can be obtained from its nearest neighbors, AR(1) landscapes are "maximally disordered," in the sense that the time series obtained by looking at the fitnesses along a nonself-intersecting walk has maximum entropy out of all statistically isotropic, Gaussian landscapes with a given correlation coefficient ρ. This fact follows immediately from the Markovian nature of the AR(1) process. We also note that the autocorrelation function of a nonself-intersecting random walk on an AR(1) landscape is given by

$$\bar{R}(s) = \rho^s = e^{-s/\lambda}.$$

This function gives the correlation between fitnesses of points s steps apart. Beyond $\lambda = 1/|\ln\rho|$ steps, fitnesses are effectively uncorrelated, so that ρ is a fundamental descriptive parameter of an AR(1) landscape.

Alternatively, one may consider the correlation $R(d)$ of points separated by a Hamming distance d. These two autocorrelation functions are related by

$$\bar{R}(s) = \sum_{d=0}^{s} \varphi_{sd} R(d)$$

where φ_{sd} denotes probability that a random walk of s steps begins and ends in two sequences with Hamming distance d (Weinberger 1990). The coefficients φ_{sd} can be obtained recursively for a free random walk on a sequence space (Fontana *et al.* 1991). For the Boolean Hypercube one finds

$$\varphi_{sd} = \frac{\nu - d + 1}{\nu}\varphi_{s-1,d-1} + \frac{d+1}{\nu}\varphi_{s-1,d+1}$$

Note that $\bar{R}(1) = R(1)$. While $\bar{R}(s)$ is much easier to obtain in numerical experiments, $R(d)$ is readily accessible for model landscapes.

As an aide to our intuition, the next two sections of this paper consider two simple model landscapes, the N-k model (Kauffman *et al.* 1988, 1989) and the p-spin model (Amitrano *et al.* 1989), both of which are roughly AR(1). Because the p-spin model has more structure than the N-k model, the deviations from an exponential autocorrelation function are more marked, but only somewhat more. We conclude that our approach is robust with respect to perturbations of the underlying model. In the fourth section, we show that these landscapes have qualitatively the same statistical properties as the model RNA landscapes introduced by Fontana and Schuster (1987), because they, too, are roughly AR(1).

With these three landscape exemplars in hand, we return in the final section of this paper to the self-similarity question. We argue that landscapes fall into qualitatively distinct classes, depending on how their correlation length scales with the overall size of the landscape. If the correlation length is proportional to the diameter of the landscape, which is, in fact the case for the N-k, p-spin, and some of the RNA landscapes, correlations exist on all length scales. Such landscapes have much in common with Brownian motion and can properly be called self-similar. In particular, AR(1) landscapes exhibit an independence of increments between neighbors that has the flavor of the independence of increments in Brownian motion.

2. The N-k Model

The N-k model assigns a real valued "fitness" to the bit string $\mathbf{b}$ by first assigning a real valued "fitness contribution", f_i , to the i^{th} bit, b_i, in $\mathbf{b}$. Each such assignment depends, not just on i and the value of b_i, but also on $0 \leq k < N$ other bits, which we call its "neighbors". The fitness contribution of each site is a random function, $f_i(\mathbf{s}_i)$, of the substring, $\mathbf{s}_i$, formed by the i^{th} bit and its k neighbors. $f_i(\mathbf{s}_i)$ is assigned by selecting an independent random variable from some distribution, such as the uniform or Gaussian distributions, for each of the 2^{k+1} possible values of $\mathbf{s}_i$, thus generating a "fitness table" for the i^{th} site. There is a different, independently generated table for each of the N sites. Then, given any string of N bits, the total fitness of the string, F, is defined as the average of the fitness contributions of each site; that is,

$$F(\mathbf{b}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{s}_i).$$

(See fig. 2 for a pictorial description of the model.) The use of a probability distribution in assigning the fitness contributions can be interpreted either as an admission of ignorance of the true nature of the complex couplings

Figure 2: Graphical representation of the N-k model for N=8, k=2, adjacent neighborhoods, showing the tables used for computing the 5th and 6th site fitnesses, but not the others, because they are computed by a similar procedure. In the 8 bit string whose fitness is to be computed, the substring consisting of bits 4, 5, and 6 is "101", so the the 5th site fitness is 0.73, as per the underlined entry in the table at left. Similarly, the substring consisting of bits 5, 6, and 7 is "010", so the 6th site fitness is 0.29. In general, there will be 2^{k+1} entries in these tables. The ith of the N tables is indexed by the k+1 bit substring formed by concatenating bit i with its k "neighbors".

between the bits or as an attempt to capture the typical statistical properties of a wide class of landscapes with k interconnections per bit.

One other aspect of the N-k model must be specified; namely, the way in which the substrings, $\mathbf{s}_i$, are chosen. The simplest — but not the only — way of choosing neighbors, at least for even k, is to use the k sites adjacent to site i; that is, the bits at sites $i - k/2$ through $i + k/2$. As in the original formulation of the model, we introduce periodic boundary conditions to assign neighbors to sites i with $i \leq k/2$ and $i \geq N - k/2$. In other words, we assume that the sites are arranged in a circle, such that site N is next to site 1. (Periodic boundary conditions are chosen because they minimize chain length dependent end effects and because we are interested in bulk properties only). This choice of neighbors gives rise to a class of short range spin glasses. Alternatively, we could assign the neighbors by randomly selecting, for each site i, k other sites, a definition that also makes sense for odd k. This assignment of neighbors makes the model similar to a long range, dilute spin glass. Rather surprisingly, local features of the landscape — the height of local optima, and the length of typical "up hill" walks through a series of fitter neighbors to these local optima — were remarkably insensitive to the details of how the $k + 1$ bit substrings were chosen in computer simulations (Kauffman *et al.* 1988, 1989).

We note that tuning k alters the ruggedness of N-k landscapes, which gives us hope that the N-k model can do a better job than the SK model in capturing the statistical properties of RNA landscapes. For $k = 0$, N-k landscapes have only one local optimum, i.e. the global optimum, uphill walks from random starts to this optimum need $\mathcal{O}(N)$ steps and a transition to a one mutant neighbor (i.e. the flip of a single bit) typically alters fitness by an amount $\mathcal{O}(1/N)$. In contrast, the N-k model for $k = N - 1$ yields Derrida's (1981) random energy landscape. As was shown recently (Kauffman & Levin 1987, Macken *et al.* 1988, 1991, Weinberger 1988, 1991a, Bak *et al.* 1992, Flyvbjerg & Lautrup 1992), such random landscapes have very many local optima ($2^N/(N+1)$, on average), walks to optima are short ($\mathcal{O}(\ln N)$), and only a small fraction of local optima are accessible from any initial string. The details of what happens for $1 \ll k \leq N$ are described in (Weinberger 1991a).

For both the adjacent and random neighbor models, the autocorrelation function has been computed exactly: for the adjacent model (Weinberger 1992),

$$R_A(d) = 1 - \frac{d(k+1)}{N} + \frac{d}{N\binom{N-1}{d-1}} \sum_{l=1}^{k} (k+1-l) \binom{N-l-1}{d-2}$$

and for the random model (Fontana *et al.* 1992),

$$R_R(d) = \left(1 - \frac{d}{N}\right) \left(1 - \frac{k}{N-1}\right)^d$$

The behavior of both of these expressions has been shown (Weinberger 1992) to be similar to that of the autocorrelation,

$$R_{PT}(d) = (1 - k/N)^d,$$

of the "P-T" model. The P-T model (Fontana *et al.* 1993) is a modification of the random neighbor N-k model in which only the bit values of the randomly chosen neighbors are used in computing the fitness tables. Henceforth, we assume $k+1$ neighbors in the P-T model for ease of comparison with the N-k model. For $k, d \ll N$, it is easily verified that

$$R_A(d) = 1 - \frac{d(k+1)}{N} + \frac{d(d-1)k(k+1)}{2N^2} + \mathcal{O}\left(\left(\frac{dk}{N}\right)^3\right),$$
$$R_{PT}(d) = 1 - \frac{d(k+1)}{N} + \frac{d(d-1)(k+1)^2}{2N^2} + \mathcal{O}\left(\left(\frac{dk}{N}\right)^3\right),$$

and

$$R_R(d) = 1 - \frac{d(k+1)}{N} + \frac{d(d-1)k(k+2)}{2N^2} + \mathcal{O}\left(\left(\frac{dk}{N}\right)^3\right).$$

It follows that all three landscapes have roughly the same statistical properties; that is, those described by the corresponding AR(1) landscape.

3. The p-Spin Model

The p-spin model outlined in this section has been proposed by Amitrano *et al.* (1989) as an alternative to the N-k model as a tunable rugged landscape. It is inspired by the use of spin glasses as models of biological fitness landscapes by Anderson (1983) and Rokhsar *et al.* (1986).

The model is defined on spin vectors, σ, with components $\sigma_i \in \{+1, -1\}$, and neighbors are those configurations separated by single spin-flips. Hence, the configuration space is isomorphic to the Boolean Hypercube. The Hamiltonian of the p-spin model is given by

$$\mathcal{H}_p(\sigma) = \sum_{i_1 < i_2 < \ldots < i_p} J_{i_1 i_2 \ldots i_p} \sigma_{i_1} \sigma_{i_2} \ldots \sigma_{i_p}$$

where the coefficients $J_{i_1 i_2 \ldots i_p}$ are independent, uncorrelated random variables. Taking $p = 2$, we recover the SK spin glass described in the first section of this paper.

Amitrano *et al.* (1988) state without proof that the correlation length of the p-spin model is $\lambda \approx \frac{N}{2p}$. Implicit in this claim is the assumption that this model is exactly AR(1). We present here an exact calculation of $R(d)$, allowing us to discuss the nature of the approximation.

Without loss of generality we may assume that the mean of the J's, $\langle J_{ij}\rangle$, is zero. Since the statistical properties of $\mathcal{H}$ do not change by either a gauge transformation — relabeling of spin states by complementation of specific bits — or a permutation of the bits, we may calculate the autocorrelation function from

$$R(d) = \frac{\langle \mathcal{H}(\mathbf{1})\mathcal{H}(\Delta)\rangle}{\mathrm{Var}[\mathcal{H}(\mathbf{1})]}$$

where

$$\mathbf{1} = (1,1,1,1,1,\ldots,1,1,1)$$

and

$$\Delta = (\underbrace{-1,-1,-1,\ldots,-1}_{d\ \text{repetitions}},\underbrace{1,1,1,\ldots,1,1}_{N-d\ \text{repetitions}}).$$

But

$$\begin{aligned}\mathrm{Var}[\mathcal{H}(\mathbf{1})] &= \sum_{\substack{i_1<i_2<\ldots<i_p\\ k_1<k_2<\ldots<k_p}} \langle J_{i_1 i_2\ldots i_p} J_{k_1 k_2 \ldots k_p}\rangle \sigma_{i_1}^2\sigma_{i_2}^2\ldots\sigma_{i_p}^2\\ &= \mathrm{Var}[J]\sum_{i_1<i_2\ldots<i_p} 1\\ &= \mathrm{Var}[J]\binom{N}{p},\end{aligned}$$

because the J's are uncorrelated. Similarly we have

$$\langle \mathcal{H}(\mathbf{1})\mathcal{H}(\Delta)\rangle = \mathrm{Var}[J]\sum_{i_1<i_2<\ldots i_p} \sigma_{i_1}\sigma_{i_2}\ldots\sigma_{i_p}$$

The indices i_k are chosen from N possibilities of which d are distinguished because they correspond to $\sigma_{i_k} = -1$. The contribution of each product of σ's to the sum is either $+1$ or -1 depending on whether the number of distinguished spins, ℓ, in the product is even or odd, respectively. The number of ways choosing the indices such that exactly ℓ of them are distinguished is

$$c_\ell = \binom{d}{\ell}\binom{N-d}{p-\ell},$$

for $\max(0, p+d-N) \leq \ell \leq \min(d,p)$. (A well known identity assures us that $\sum_\ell c_\ell = \binom{N}{p}$, as must be the case.) We conclude that

$$\begin{aligned}\langle \mathcal{H}(\mathbf{1})\mathcal{H}(\Delta)\rangle &= \mathrm{Var}[J]\Big\{ \sum_{\ell\, even}^{*} \binom{d}{\ell}\binom{N-d}{p-\ell} - \sum_{\ell\,\mathrm{odd}}^{*} \binom{d}{\ell}\binom{N-d}{p-\ell}\Big\} \\ &= \mathrm{Var}[J]\Big\{ \binom{N}{p} - 2\sum_{\ell\,\mathrm{odd}}^{*} \binom{d}{\ell}\binom{N-d}{p-\ell}\Big\}\end{aligned}$$

and therefore

$$R_p(d) = 1 - \frac{2}{\binom{N}{p}} \sum_{\ell\,\mathrm{odd}}^{*} \binom{d}{\ell}\binom{N-d}{p-\ell},$$

where the asterix in all of the above sums reflects the condition on ℓ imposed by the above definition. For $d \cdot p \ll N$,

$$R_p(d) \approx 1 - \frac{2dp}{N} + \frac{2d(d-1)p(p-1)}{N^2} + \mathcal{O}(N^{-3}),$$

suggesting that the p-spin landscape should have similar properties to the AR(1) landscape with

$$\bar{R}(s) = \Big(1 - \frac{2p}{N}\Big)^s,$$

from which the estimate $\lambda \approx \frac{N}{2p}$ follows. If p is proportional to N, the resemblance to the corresponding AR(1) landscape is even more striking, because we then have

$$R(d) = \Big(1 - \frac{2p}{N}\Big)^d \Big(1 + \mathcal{O}\big(\max(\frac{1}{p}, \frac{1}{N-p})\big)\Big).$$

For the N-k-model the characteristic length is given by (Weinberger 1990)

$$\lambda_{Nk} \approx \frac{N}{k+1}.$$

Hence, if $k+1 \approx 2p$, the corresponding N-k and p-spin landscapes should have roughly the same correlation structure. When $p = 2$ (the SK spin glass), we have

$$R_2(d) = 1 - 4\frac{d}{N-1} + 4\frac{d^2}{N(N-1)}.$$

For the SK model the autocorrelation function $\bar{R}(s)$ along a random walk can be derived using the techniques derived by Stadler & Happel (1992). Denote $\Delta_m(s) = \sum_d \varphi_{sd} d^m$. Using the recursion for the coefficients (*cf.* sect. 2) it is straight forward to derive recursions for $\Delta_m(s)$

$$\Delta_1(s) = 1+(1+2/N)\Delta_1(s-1) \qquad \Delta_2(s) = 1+2\Delta_1(s-1)+(1-4/N)\Delta_2(s-1)$$

The autocorrelation function along the walk is

$$\bar{R}_2(s) = \sum_{d=0}^{s} \varphi_{sd} R_2(d) = 1 - \frac{4}{N-1}\Delta_1(s) + \frac{4}{N(N-1)}\Delta_2(s).$$

Substituting the recursions for $\Delta_m(s)$ and some algebra yields

$$\begin{aligned} \bar{R}_2(s) &= \left(1 - \frac{4}{N}\right)\left(1 - \frac{4}{N-1}\Delta_1(s-1) + \frac{4}{N(N-1)}\Delta_2(s-1)\right) = \\ &= (1 - \frac{4}{N})R(s-1) \end{aligned}$$

Hence the SK model is strictly AR(1) with autocorrelation functions

$$\bar{R}_2(s) = \left(1 - \frac{4}{N}\right)^s.$$

4. Statistical properties of RNA landscapes

RNA sequences are encoded as strings of length N over the alphabet $\{$**G,C,A,U**$\}$. The canonical distance measure between two sequences of equal length N is the Hamming distance. The physico-chemical properties of RNA molecules, and, in particular, their free energies of folding, ΔG, depend crucially on the spatial structure of the molecule.

In the case of RNA the process of folding the 1D primary structure (sequence) into the 3D tertiary structure can be decomposed into two steps:

1) Folding of the sequence into a planar secondary structure by formation of complementary base pairs **G≡C** and **A=U**. (The weaker **G−U** pairs are also considered in this step.)

2) 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. The dominant role of the secondary structure is well documented in nature by the conservation of secondary structure elements in evolution (Cech 1988).

At present there are no reliable models for the prediction of tertiary structure, whereas efficient algorithms exists for the prediction of the secondary structure (Zucker 1984, McCaskill 1990). The assumption built into these algorithms through the definition of the secondary structure are:

1) Secondary structures are planar graphs, i.e. there are no knots.

2) Pseudo-knots and triple-helices are considered as part of the tertiary structure.

3) Secondary structures can be partioned into elements contributing additively to the thermodynamic properties. All non-additive contributions are attributed to the tertiary structure.

The structural elements of secondary structures are

1) *stems* or *stacks*, the double-helical regions,

2) *loops* and *bulges* consisting of internal unpaired bases,

3) *joints*, which are stretches of unpaired digits joining freely movable substructures, and

4) *free ends*.

It is this additivity of the contribution of the structural elements, which suggest *a priori* a similarity of the RNA folding landscapes with the much simpler p-spin and N-k models.

Fontana and co-workers (1987, 1988) proposed simple estimates of replication rate and degradation rates of RNA *in vitro* based on the following principles:

1) Replication operates on single strands (Biebricher *et al.* 1984).
2) Unzippering of helices is cooperative (Pörschke 1977).
3) The hydrolytic degradation reaction is not cooperative.
4) As a first approximation, sequence specificity is neglected, except to the degree that it determines secondary structure, so that rate constants depend only on secondary structure.

The replication and degradation estimates themselves are given by

$$A(I_k) = \kappa_r + \kappa_1 \sum_j \frac{n_j(I_k)(1 - n_j(I_k))^3}{(1 + n_j(I_k))^4 + L}$$

$$D(I_k) = \kappa_d + \kappa_2 \sum_j \frac{u_j(I_k)}{u_m} \exp\{(u_j(k) - u_m)/u_m\} + \kappa_3 \sum_l \frac{N_l(I_k)}{N},$$

respectively. $n_j(I_k)$, $u_j(I_k)$ and $N_j(I_k)$ denote the number of base-pairs in stack j of sequence I_k, the number of nucleotides in loop j and the number of unpaired digits in joints and free ends, respectively. The remaining symbols are constants. For details see Fontana & Schuster (1987). The effective replication rate $E(I_k)$ is obtained as the difference of replication and degradation rate.

Extensive numerical studies of the free energy landscape and the three kinetic landscapes have recently been performed (Fontana *et al.* 1991) for both the full **AUGC**-landscape, i.e. for molecules composed of all four bases, and the restriction to **GC**-only sequences of given length N. The configuration space of the **GC**-only space is isomorphic to the Boolean Hypercube and thus with the configuration spaces of both the Nk-model and the p-spin models.

The statistical properties of these 8 landscapes

$$I \mapsto \Delta G(I)$$

$$I \mapsto A(I)$$

$$I \mapsto D(I)$$

$$I \mapsto E(I) = A(I) - D(I)$$

for both the full and the restricted alphabet can be summarized as follows (Fontana *et al.* 1991, 1992, 1993)

- The ΔG-landscapes are nearly AR(1) for both the full and the restricted alphabet
- The autocorrelation function of the kinetic landscapes deviate significantly from a single decaying exponential. The deviations are small enough, however, to define an approximate "characteristic length-scale" λ by $R(\lambda) = e^{-1}$.
- The correlation length of the free energy landscapes scales linearly with the sequence length N for both alphabets.
- The kinetic constants do not scale with N in **GC**-space, but appear to be bounded above.
- In the **GCAU** case the correlation length of the functions E and D seem to scale linearly with N, whereas A probably saturates.
- There is a partial, but not exact, symmetry between the secondary structures of two complementary RNA strands. Because similar structures usually exhibit similar kinetic and thermodynamic properties, and because the only sequences separated by a distance N are complementary, assuming the **GC**-only alphabet, we expect $R(N) \gg 0$. Complementarity plays a much smaller role in **GCAU**-case.

Table 1 suggests a classification into two qualitatively different families of value landscapes:

I landscapes with long range correlations ($\lambda = \mathcal{O}(N)$).

II landscapes with only short range correlations (($\lambda = \mathcal{O}(1)$.)

Table 1: Characteristic length-scales of the correlation in RNA landscapes taken from Fontana *et al.* (1991). All data refer to $N \leq 100$.

Landscape	**GC**	**GCAU**
ΔG	$\sim N/12$	$\sim 0.3N$
A	≤ 3	≤ 15
D	≤ 3.5	$\sim N/4$
E	≤ 4	$\sim N/5$

It also explains why Eigen *et al.* (1989) were unable to fit RNA kinetic data to an SK landscape: the two landscapes have significantly different statistical structures.

5. Theory of Random Fractal Landscapes

In this section, we discuss the connection between landscape correlation and self-similarity (i.e. "fractalness"), thus establishing the precise meaning of this last term. We conclude with some comments on the constraints that our results place on models of RNA fitness landscapes.

We start with a definition of a random fractal landscape taken directly from Sorkin (1988), who built on the idea of fractional Brownian motion (Voss 1986). In order to make the present paper accessible to a wider audience, we include a motivation of this definition and a discussion of some relevant technical details in an Appendix.

Definition. *A landscape is fractal if*

$$\langle \|X_i - X_j\|^2 \rangle \propto d^{2h}(i,j),$$

where the two configurations i and j with normally distributed fitnesses X_i and X_j, respectively, are separated by a distance $d(i,j)$.

In other words, the variance of the difference between the fitnesses of two configurations scales as a power law with their distance from each other. As a practical matter, the distance between configurations for any finite configuration space is bounded, so that Sorkin's definition makes sense only for distances small compared to this upper bound. We call this bound the *diameter* of the configuration space, which we denote diam $\mathcal{C}$.

Sorkin's definition can be reformulated as

$$1 - R(d) \propto d^{2h}.$$

For distances d small compared to the correlation length λ, all of the landscapes described above, as well as every other AR(1) landscape, satisfy

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

and therefore approximately satisfy Sorkin's definition with $h = 1/2$.

It is useful to scale the distance d by the diameter (the maximum distance) of the configuration space:

$$\xi = d/\mathrm{diam}\ \mathcal{C}$$

The rescaled autocorrelation function may be written as

$$r(\xi) = 1 - \frac{\mathrm{diam}\ \mathcal{C}}{\lambda}\xi + \mathcal{O}(\xi^2)$$

To the degree that the linear approximation holds, Sorkin's definition is satisfied. More generally, this rescaling leads to the following classification of landscapes by their behavior in the limit diam $\mathcal{C} \to \infty$:

I $\lim\limits_{\mathrm{diam}\ \mathcal{C}\to\infty} r(\xi) = 1 - \alpha\xi + \mathcal{O}(\xi^2)$, for $\alpha = \mathcal{O}(1)$. Examples, apart from those discussed above, include the landscape of the travelling salesman problem (TSP) with transpositions or 2-opt moves (Stadler & Schnabl 1992) the landscape of the graph bipartioning problem (Stadler & Happel 1992), the Hopfield-Hamiltonian (Tarazona 1992). the graph matching problem, or the low autocorrelated binary string problem

(Stadler 1992). Class **I** landscapes are "trivial" random fractals with $h = \frac{1}{2}$.

II $\lim_{\text{diam } \mathcal{C} \to \infty} r(\xi)$ is discontinuous for $\xi \to 0$. This corresponds to "extremely rugged" landscapes. One may divide them into two subclasses:

IIa $\lim_{\text{diam } \mathcal{C} \to \infty} r(\xi) = \delta(\xi, 0)$ (or more generally consists of a finite sum of delta functions). One may define two subclasses depending on whether the unscaled correlation length diverges as diam $\mathcal{C} \to \infty$ or whether it is uniformly bounded from above.

An example of the first subclass is the TSP with canonical transpositions defining the neighborhood. Let N be the number of cities. The correlation length is $\mathcal{O}(N)$, whereas the diameter of the landscape is $\mathcal{O}(N^2)$. See below for details.

Examples of the second subclass are probably the kinetic landscapes for the **GC**-alphabet and of course the totally uncorrelated landscape of the random energy model (Derrida 1981).

IIb $\lim_{\text{diam } \mathcal{C} \to \infty} r(\xi)$ does not vanish in a neighbourhood U of $\xi = 0$ but the autocorrelation function is bounded from above by $\gamma < 1$ for $\xi \in U \setminus \{0\}$. An example is provided by the asymmetric TSP with 2-opt moves.

III $\lim_{\text{diam } \mathcal{C} \to \infty} r(\xi) = 1 - \beta\xi^2 + o(\xi^2)$, for $\beta = \mathcal{O}(1)$. In contrast to all previous cases, the second derivative of the autocorrelation function, $r''(s)$ exists for all s and is continuous for $s = 0$. Note that it is only because the linear term in ξ vanishes that we can satisfy both the symmetry condition $r(s) = r(-s)$ imposed by stationarity and the continuity condition $r(0^+) = r(0^-)$. These properties of r guarantee that the corresponding nonself-intersecting walk is, in the infinite diameter limit, a mean square differentiable stochastic process (Papoulis 1965).

Under mild additional conditions, such processes are differentiable in the ordinary sense. The corresponding landscapes form the second class of trivial fractals, corresponding formally to Sorkin's definition of fractals with $h = 1$. These are the smoothest possible landscapes, since any function $f(x) = 1 - \mathcal{O}(x^{2+q})$ with $q > 0$ cannot be autocorrelation function of a stochastic process. A simulation of evolutionary processes with a smooth fitness function (Galar 1991) mimics such landscapes.

IV $\lim_{\text{diam }\mathcal{C}\to\infty} r(\xi) = 1 - c_p\xi^p + o(\xi^p)$ with $p \neq 1, 2$, and thus correspond to non-trivial random fractals. Such landscapes have been constructed by Sorkin (1988) but have not (yet) been found in "real life" examples. Because the corresponding nonself-intersecting random walks would not be Markovian and would therefore violate the maximum entropy condition discussed in the introduction, we conjecture that they will rarely, if ever, be found.

We emphasize that, in the same approximation in which class **I** landscapes are fractal, they are also AR(1):

$$(1-q)^{\xi} = 1 - q\xi + \mathcal{O}(\xi^2)$$

One may express this as: "An AR(1) landscape is fractal iff the correlation length is proportional to the diameter of the landscape". If this is true, there are correlations on all length scales, and, as is the case with Brownian motion, there is no preferred length scale.

Note also that the classification scheme presented above applies to landscape geometries as different from the hypercube as the configuration space of the TSP, as we now show in more detail. The correlation length for the N city TSP is approximately $\frac{N}{b}$ where b is the average number of bonds broken by a single move (Stadler & Schnabl 1992). A canonical transposition is the interchange of two neighbouring cities along the tour (i.e. to cycles of the form $(i, i+1)$, as opposed to the full set of possible transpositions of two cities); therefore, $b = 2$ for symmetric and $b = 3$ for asymmetric distances between the cities. The diameter of the configuration space with respect to this move set is the minimum number of canonical transpositions necessary to invert a cyclic permutation,

$$\mathrm{diam}\mathcal{C} = (N-1) + (N-2) + ... + 1 = \frac{1}{2}N(N-1)$$

since the last entry has be exchanged $N-1$ times, the next one $N-2$ times, etc. Therefore the scaled correlation length

$$\frac{\lambda}{\mathrm{diam}\mathcal{C}} = \frac{b}{2(N-1)} \to 0$$

vanishes for large N, and we have a class **II** landscape. For 2-opt moves and the full set of transpositions the diameter of the landscape is bounded from above by N and from below by $N/2$, thus $\lambda/\mathrm{diam}\mathcal{C}$ converges towards a finite number. These are fractal class **I** landscapes.

For evolutionary modelling in general and RNA landscapes in particular, the main conclusion is that model landscapes must be no smoother than class I landscapes, which we expect to be typical. Indeed, previous sections show that λ is determined by a fixed number of interactions per site (i.e. k in the N-k model, or p in the p-spin model), divided by the diameter of the landscape. However, some kinetic properties of biological macromolecules seem to be even more sensitive to changes in the sequences, so that there might well be kinetic landscapes which are as rugged as the model proposed by Fontana *et al.* (1991, 1992, 1993) for **GC** sequences. Recent results by Tacker *et al.* (1992) suggest that landscapes of rate constants generically are class **II** landscapes. In any case, landscapes appropriate for evolutionary modelling should probably be considerably more rugged than the spin glass landscapes used by Anderson (1983) and others (see e.g. Mézard *et al.* 1987).

In addition to its role in characterizing the statistics of landscapes, and thus the dynamics of evolutionary optimization on them, the correlation length of a landscape reflects the computational difficulty involved in finding its global optima. The qualitative difference between relatively uncorrelated RNA folding landscapes and previously studied, more highly correlated landscapes suggests that there are at least two classes of landscape optimization problems, and that, perhaps, different strategies should be employed for each class.

Acknowledgments

Stimulating discussions with Manfred Eigen, Peter Schuster, Pedro Tarazona and Subbiah Baskaran are gratefully acknowledged.

References

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

[2] Anderson P.W. (1983). Suggested Model for Prebiotic Evolution: The Use of Chaos. *PNAS* **80**, 3386–3390.

[3] Bak P., H. Flyvbjerg, B. Lautrup. (1992). Coevolution in a rugged fitness landscape. *Phys. Rev. A[15]* **46**, 6714–6723.

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

[5] Breiman, L. (1986). *Probability.* Addison-Wesley, Reading, Ma.

[6] Cech, T.R. (1988). Conserved Sequences and Structures of group I introns: building an active site for RNA catalysis. *Gene* **73** 259–271.

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

[8] Eigen, M. (1971). Selforganization of Matter and the Evolution of Biologocal Macromolecules. *Naturwissenschaften* **58**, 465–523.

[9] Eigen, M. (1985). Macromolecular evolution: dynamical ordering in s equence space, in *Emerging Synthesis in Science. Proceedings of the Founding Workshops of the Santa Fe Institute*, D. Pines (ed.), pp. 25-69. The Santa Fe Institute, Santa Fe, New Mexico.

[10] Eigen, M. and P. Schuster (1979) *The Hypercyle.* Springer.

[11] Eigen, M., J. McCaskill and P. Schuster (1989). The Molecular Quasispecies. *Advances in Chemical Physics* **75**, 149–263.

[12] Fontana W. and P. Schuster. (1987). A computer model of evolutionary optimization. *Biophysical Chemistry* **26** 123-147.

[13] Fontana W., W. Schnabl, and P. Schuster. (1988). Physical aspects of evolutionary optimization and adaptation. *Phys. Rev. A* **40**, 3301–3321.

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

[15] Fontana W., D.A.M. Konings, P.F. Stadler, and P. Schuster. (1992). Statistics of RNA Secondary Structures. *Biopolymers* in press.

[16] Fontana W., P.F. Stadler, E. Bauer, T. Griesmacher, I.L. Hofacker, M. Tacker, P. Tarazona, E.D. Weinberger, and P. Schuster. (1993). Statistical Properties of RNA free energy landscapes. *Phys. Rev. E* in press.

[17] Flyvbjerg H. and B. Lautrup. (1992). Evolution in a rugged fitness landscape. *Phys. Rev. A[15]* **46**, 6724–6730.

[18] Galar R. (1991). Simulation of local evolutionary dynamics of small populations. *Biol. Cybern.* **65**, 37-45.

[19] Kauffman, S. and S. Levin (1987). Towards a General Theory of Adaptive Walks on Rugged Landscapes. *J. Theor. Biol.* **128**, 11–45.

[20] Kauffman, S. and E. Weinberger (1989). The N-k Model of Rugged Fitness Landscapes and Its Application to Maturation of the Immune Response, *J. Theor. Biol.* **141** 211–245.

[21] Kauffman, S., Weinberger, E., and Perelson, A. (1988). 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.

[22] Macken, C.A. and A.S. Perelson. (1989). Protein Evolution on Rugged Landscapes. *Proc. Natl. Acad. Sci. (USA)* **86**, 6191–6195.

[23] Macken, C.A., P.S. Hagan and A.S. Perelson. (1991). Evolutionary Walks on Rugged Landscapes. *SIAM J. Appl. Math.* **51**, 779–827.

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

[25] McCaskill, J. (1990) The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structures. *Biopolymers* **29**, 1105–1119.

[26] Mézard M., G. Parisi and M.A. Virasoro. (1987) *Spin Glass Theory and Beyond.* World Scientific, Singapore.

[27] Papoulis, A. (1965). *Probablity, Random Variables, and Stochastic Processes*, M^{c}Graw-Hill, New York.

[28] Pörschke, D. (1977). Elementary steps of base recognition and helix–coil transitions in nucleic acids. In: I. Pecht and R. Rigler (eds.)

Chemical Relaxation in Molecular Biology. Springer-Verlag, Berlin pp.191-218.

[29] Rokhsar, D.S., P.W. Anderson and D.L. Stein. (1986) Selforganization in prebiological systems: simulations of a model of the origin of genetic information. *J. Mol. Evol.* **23**, 119-126.

[30] Schuster P. (1986). The Physical Basis of Molecular Evolution. *Chem. Scripta* **26B**, 27–41.

[31] Schuster P. (1987). Structure and Dynamics of Replication-Mutation Systems. *Phys. Scripta* **35**, 402–416.

[32] Sherrington, D. and S. Kirkpatrick. (1975) Solvable Model of a Spin Glass. *Phys. Rev. Lett.* **35**, 1792-1796.

[33] Sorkin, G. (1988). Combinatorial Optimization, Simulated Annealing, and Fractals, IBM Research Report RC13674 (No. 61253).

[34] Stadler, P.F. and W. Schnabl (1992). The Landscape of the Travelling Salesman Problem, *Physics Letters A*, **161**, 337-344.

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

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

[37] Tacker M., W. Fontana, P.F. Stadler, and P. Schuster. (1992) Statistics of RNA Melting Kinetics. preprint.

[38] Tarazona, P. (1992) Error Thresholds for Molecular Quasispecies as Phase Transitions: From Simple Landscapes to Spin-Glass Models. *Phys. Rev. A[15]* **45**, 6038–6050.

[39] Voss, R. (1986). Characterization and Measurement of Random Fractals, *Physical Scripta* **T13**, 257–260.

[40] Weinberger, E. (1988). A more rigorous derivation of some properties of uncorrelated fitness landscapes. *J. Theor. Biol.* **134**, 125–129.

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

[42] Weinberger, E. (1991a). Local Properties of Kauffman's N-k model, a Tuneably Rugged Energy Landscape, *Physical Review A*, **44**, 6399-6413.

[43] Weinberger, E. (1991b). Fourier and Taylor Series on Fitness Landscapes, *Biological Cybernetics*, **65**, 321-330.

[44] Weinberger, E. (1991c). Measuring Correlations in Energy Landscapes and Why It Matters, in: *Information Dynamics*, Proceedings of the NATO Advanced Study Institute on Information Dynamics, June 1990, Kaufbeuren, Bavaria; H. Atmanspacher, H. Scheingraber, eds., Plenum Publishing, New York.

[45] Weinberger, E. (1992). NP-Completeness of Kauffman's N-k Model, A Tunably Rugged Fitness Landscape. Submitted to *Acta Appl. Math.*

[46] Wright, S. (1932), The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: *Proceedings 6th Congress on Genetics* Vol.1, p.356.

[47] Zuker M. and D. Sankoff. (1984). RNA Secondary Structures and their Prediction. *Bull. Math. Biol.* **46**, 591–621.

Appendix: Fractals and Fractional Brownian Motion

The purpose of this section is to motivate the definition of a random fractal landscape used in the text, and to clarify the relationship between random fractal landscapes and Brownian motion. We first explain exactly what we mean by the buzz-words "fractal" and "self-similar": A set of points $S \subset \mathbb{R}^N$ is self-similar (or fractal) if it is the disjoint union of a finite number of copies, possibly translated and/or rotated, of the set $rS = \{r\mathbf{x}, \mathbf{x} \in S\}$, where $r\mathbf{x}$ is the usual multiplication of the vector, $\mathbf{x}$, by the scalar r. In other words, S is self-similar if it is made up of a number of copies of itself, scaled down by a factor r, which themselves are made up of copies of themselves, again scaled down by a factor r an so on. A simple example is the Koch curve shown in Figure 3. In particular, if S is the "area under the curve" of some scalar function $X(t)$ on the interval $0 < t \leq 1$ — that is, the region of $\mathbb{R}^2$ bounded by $t = 0$, $t = a$, $X = 0$, and $X = X(t)$ — the concatenation of C copies of rS is the graph under the curve $rU(t)$, $0 < t \leq C$, where

$$U(t) = \begin{cases} X(t) & \text{for } 0 < t \leq a, \\ X(t-a) & \text{for } a < t \leq 2a, \\ \vdots & \\ X(t-(C-1)a) & \text{for } (C-1)a < t \leq Ca. \end{cases}$$

If S is self-similar, there will exist values of r and C such that $rU(Ct) = X(t)$. Furthermore, as $a \to \infty$, $rU(Ct) \to rX(Ct)$ over the whole real line, thus motivating the definition of a self-similar function as one that satisfies the functional equation $rX(Ct) = X(t)$, for some values of the parameters r and C. Note that the above construction is still possible and the definition still makes sense even if C is not an integer.

More generally, we may even replace the deterministic function X by a random process. Suppose now that $X(t)$ and $rX(Ct)$ are not exactly identical, but they *are* identical random processes. $X(t)$ is then said to be "a statistically self-affine fractal" or a "random fractal." It is this generalization that gives these ideas their power: the original and rescaled functions are rarely exactly identical in applications, but are often identical in a statistical sense. Random fractals arise when there is no "natural" scaling of the

Figure 3: Koch's curve is an example of a fractal curve in the plane. Its is constructed by the following iteration scheme starting with the unit interval: each segment is partitioned into 3 pieces and the middle one is replaced by a $\wedge$-shaped spike the length of the edges of which equal the length of the replaced segment. The fractal dimension of this curve is thus $\log 4/\log 3$.

independent variable, as, for example, in Brownian motion, which we denote $X_{1/2}(t)$, to emphasize its dependence on the scaling parameter defined below.

Brownian motion is characterized mathematically as a continuous, stationary Gaussian random process whose increments,

$$\Delta_{1/2}(|t_2 - t_1|) = X_{1/2}(t_2) - X_{1/2}(t_1),$$

satisfy

(i) $\langle \Delta_{1/2}(|t_2 - t_1|) \rangle = 0$.

(ii) $\mathrm{Var}\{\Delta_{1/2}(|t_2 - t_1|)\} = D(|t_2 - t_1|)$, for some constant $D > 0$.

The probability that $\Delta_{1/2}(|t_2 - t_1|)$ assumes the particular value x for a given time increment $|t_2 - t_1|$ is

$$\Pr\{x \le \Delta_{1/2}(|t_2 - t_1|) \le x + dx\} = \frac{e^{-x^2/2D|t_2-t_1|}}{2\pi D|t_2 - t_1|}\, dx,$$

from which it is clear that

$$\Pr\{\frac{x}{\sqrt{\alpha}} \le \Delta_{1/2}(|t_2 - t_1|) \le \frac{x + dx}{\sqrt{\alpha}}\} = \Pr\{x \le \Delta_{1/2}(\alpha|t_2 - t_1|) \le x + dx\}.$$

In other words, the random variable $\Delta_{1/2}(\alpha|t_2 - t_1|)/\sqrt{\alpha}$ has the same probability distribution as the random variable $\Delta_{1/2}(|t_2 - t_1|)$. We write

$$\Delta_{1/2}(\alpha|t_2 - t_1|)/\sqrt{\alpha} \simeq \Delta_{1/2}(|t_2 - t_1|)$$

to emphasize that it is not the actual realizations of the two stochastic processes, but rather their probability distributions that are equal.

So-called fractional Brownian motion, $X_h(t)$, is a generalization of Brownian motion process, $X_{1/2}(t)$, that retains self-similarity. More specifically, Postulate (ii) above is replaced by

(ii') $D|t_2 - t_1|^{2h} = \mathrm{Var}\{\Delta_h(|t_2 - t_1|)\}$, for some constant h with $0 < h < 1$. Because

$$\Pr\{x \le \Delta_h(|t_2 - t_1|) \le x + dx\} = \frac{e^{-x^2/2D|t_2-t_1|^{2h}}}{2\pi D|t_2 - t_1|^{2h}}\, dx,$$

we have

$$\Delta_h(|t_2 - t_1|) \simeq \alpha^{-h}\Delta_h(\alpha t),$$

recovering the usual Brownian motion when $h = 1/2$.

It was claimed in Section 5 that successive increments of $X_h(t)$ are correlated if $h \neq 1/2$. To see this, we write, for $t_1 < t < t_2$,

$$\begin{aligned} D|t_2 - t_1|^{2h} &= \left\langle \Delta_h(t_2, t_1)^2 \right\rangle \\ &= \left\langle [\Delta_h(t_2, t) + \Delta_h(t, t_1)]^2 \right\rangle \\ &= D|t_2 - t|^{2h} + D|t - t_1|^{2h} + 2\left\langle \Delta_h(t_2, t)\Delta_h(t, t_1) \right\rangle. \end{aligned}$$

Hence,

$$\langle \Delta_h(t_2, t)\Delta_h(t, t_1)\rangle = \frac{D}{2}\left[|t_2 - t_1|^{2h} - D|t_2 - t|^{2h} - D|t - t_1|^{2h}\right],$$

from which it follows that the two increments $\Delta_h(t, t_1)$ and $\Delta_h(t_2, t)$ are negatively correlated for $h < 1/2$ and positively correlated for $h > 1/2$.

The idea of fractional Brownian motion can be generalized, first to multi-dimensional Euclidian spaces, and then to high dimensional discrete spaces. The generalization to multi-dimensional Euclidian spaces is straightforward, provided we understand that the argument of the Gaussian random function X_h is now an M-dimensional vector $\mathbf{z}$, rather than the scalar, t. M dimensional fractional Brownian "motion" is thus a surface in $\mathbb{R}^{M+1}$, the set of all $M+1$-tuples $(\mathbf{z}, X_h(\mathbf{z}))$, where

$$\left\langle |X_h(\mathbf{z}_1) - X_h(\mathbf{z}_2)|^2 \right\rangle \propto \|\mathbf{z}_1 - \mathbf{z}_2\|^{2h}.$$

Cutting the surface $X_h(\mathbf{z})$ along the direction $\mathbf{z}_0$ (i.e. a vertical plane) yields

$$\begin{aligned}\left\langle |X_h(t_2\mathbf{z}_0) - X_h(t_1\mathbf{z}_0)|^2 \right\rangle &\propto \|t_2\mathbf{z}_0 - t_1\mathbf{z}_0\|^{2h} \\ &\propto |t_2 - t_1|^{2h}.\end{aligned}$$

In other words, a multi-dimensional fractional Brownian motion cut by a plane is the scalar version of fractional Brownian motion discussed above.

It was this observation that motivated Sorkin to consider what happens when the cut is, itself, a fractal surface. More precisely, Sorkin replaced $\mathbf{z}(t) = t\mathbf{z}_0$ by a "random walk" (or "regular" Brownian motion). He was then able to show that a random walk through the multi-dimensional Euclidian fractal $X_h(\mathbf{z})$ is, itself, a random fractal with scaling parameter $h' = h/2$. Hence Sorkin's definition, given in Section 5, that

A (possibly discrete) landscape is fractal if

$$\left\langle \|X_i - X_j\|^2 \right\rangle \propto d^{2h}(i,j),$$

where the two configurations i and j with normally distributed fitnesses X_i and X_j, respectively, are separated by a distance $d(i,j)$.

Figure Captions

Figure 1. Sequence Spaces for binary sequences of length $N = 3$ and $N = 5$. All pairs of sequences with Hamming distance 1 are connected by a line. The graph defined in this way is a hypercube of dimensions N. Boolean Hypercubes are the configuration spaces for the N-k model, the p-spin model and the **GC**-only RNA model.

Figure 2. Graphical representation of the N-k model for $N = 8$, $k = 2$, adjacent neighborhoods, showing the tables used for computing the 5^{th} and 6^{th} site fitnesses, but not the others, because they are computed by a similar procedure. In the 8 bit string whose fitness is to be computed, the substring consisting of bits 4, 5, and 6 is "101", so the the 5^{th} site fitness is 0.73, as per the underlined entry in the table at left. Similarly, the substring consisting of bits 5, 6, and 7 is "010", so the 6^{th} site fitness is 0.29. In general, there will be 2^{k+1} entries in these tables. The i^{th} of the N tables is indexed by the $k + 1$ bit substring formed by concatenating bit i with its k "neighbors".

Figure 3. Koch's curve is an example of a fractal curve in the plane. Its is constructed by the following iteration scheme starting with the unit interval: each segment is partitioned into 3 pieces and the middle one is replaced by a $\wedge$-shaped spike the length of the edges of which equal the length of the replaced segment. The fractal dimension of this curve is thus $\log 4/\log 3$.