RNA Folding and Combinatory Landscapes

By

Walter Fontana[c,d], Peter F. Stadler[a,b,c], Erich G. Bornberg–Bauer[a], Thomas Griesmacher[a], Ivo L. Hofacker[a], Manfred Tacker[a], Pedro Tarazona[a,f], Edward D. Weinberger[b], and Peter Schuster[a,c,e,*]

[a] Institut für Theoretische Chemie, Universität Wien
[b] Max-Planck-Institut für Biophysikalische Chemie, Göttingen
[c] Santa Fe Institute, Santa Fe
[d] Theoretical Division, Los Alamos National Laboratory
[e] Institut für Molekulare Biotechnologie, Jena
[f] Dept. de Fisica de la Materia Condensada, Univ. Autonoma de Madrid

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

Abstract

In this paper we view the folding of polynucleotide (RNA) sequences as a map that assigns to each sequence a minimum free energy pattern of base pairings, known as secondary structure. Considering only the free energy leads to an energy landscape over the sequence space. Taking into account structure generates a less visualizable non-scalar "landscape", where a sequence space is mapped into a space of discrete "shapes". We investigate the statistical features of both types of landscapes by computing autocorrelation functions, as well as distributions of energy and structure distances, as a function of distance in sequence space. RNA folding is characterized by very short structure correlation lengths compared to the diameter of the sequence space. The correlation lengths depend strongly on the size and the pairing rules of the underlying nucleotide alphabet. Our data suggest that almost every minimum free energy structure is found within a small neighborhood of any random sequence.

The interest in such landscape results from the fact that they govern natural and artificial processes of optimization by mutation and selection. Simple statistical model landscapes, like Kauffman's n-k model, are often used as a proxy for understanding realistic landscapes, like those induced by RNA folding. We make a detailed comparison between the energy landscapes derived from RNA folding and those obtained from the n-k model. We derive autocorrelation functions for several variants of the n-k model, and briefly summarize work on its fine structure. The comparison leads to an estimate for $k = 7$ to 8, independent of n, where n is the chain length. While the scaling behaviors agree, the fine structure is considerably different in the two cases. The reason is seen to be the extremely high frequency of neutral neighbors, that is: neighbors with identical energy (and structure), in the RNA case.

I Combinatory Maps

Processes like combinatorial optimization and evolutionary adaptation take place on landscapes that result from mapping microconfigurations to energies or non-scalar entities like structures. A classic example from physics is a Hamiltonian that assigns an energy value to a spin configuration. Another instance, taken from operations research, is a geography that maps tours through a set of cities into transport costs. Many properties of those processes reflect the local and global statistical features of the landscape on which they occur. This leads to the problem of understanding what these features are, and how to study them. In this paper we make an attempt by deriving new results and integrating them with established ones.

As an example we study the biologically important landscape induced by the folding of polynucleotides (RNA). Here the microconfigurations are sequences over an alphabet of nucleotides, and scalar properties are free energies of secondary or tertiary structure formation. Another scalar could be the rate constant of a reaction involving that structure (e.g. replication of viral RNA). Non-scalar properties like the secondary structure or the 3D structure are of particular interest. Here we consider both the energy as well as the structure landscapes induced by RNA folding, and compare the former with a simple parametrized model landscape, known as the n-k model [1].

Common to all these examples is a function whose domain is a set of combinatorial complexity — where the elements represent combinations or variations of some kind — and whose range is either $\mathbb{R}$, or another set of combinatorial complexity (suitably discretized structures, for example). This motivates the following definition of a combinatory map (CM).

Definition. A CM is a quintuple $(\mathcal{X}, d_{\mathcal{X}}; \mathcal{Y}, d_{\mathcal{Y}}; f)$ where $\mathcal{X}$ and $\mathcal{Y}$ are sets endowed with metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, respectively, and f is a map $\mathcal{X} \longrightarrow \mathcal{Y}$. If

$\mathcal{Y} = \mathbb{R}$ and $d_{\mathcal{Y}}(a,b) = |a-b|$, we refer to the quintuple as a combinatory landscape (CL).

$(\mathcal{X}, d_{\mathcal{X}})$ is known as configuration space. The natural metric is induced by some physically meaningful set of operations that interconvert configurations. In the present case of RNAs a configuration (or sequence) space consists of all sequences of fixed length n over an alphabet of size κ (typically $\kappa = 4$), the interconversion operations are point mutations, and the Hamming metric provides a distance measure between sequences.

The basic problem we are concerned with here is how to investigate the major statistical features of CMs. One approach studies CMs from the point of view of a random walker [2,3]. This essentially converts spatial information into time series that can be characterized, for example, by autocorrelation functions. Another approach attempts to devise tools that reflect the statistical features of CMs as a whole. Some of these features cannot be accessed with random walks alone.

In section 2 we generalize the autocorrelation function to the case of combinatory maps, and define density surfaces as an instance for the second approach above. Section 3 gives a brief classification of landscapes by autocorrelation, section 4 presents a study of the RNA case, section 5 introduces Kauffman's n-k model, derives landscape autocorrelation functions, and briefly reviews analytic results on gradient and adaptive walks. Section 6 compares the RNA energy landscape with the n-k model. Section 7 concludes the paper.

II Autocorrelation and Density Surfaces

In the case of landscapes a random walk $\{x_1, x_2, \ldots\}$ on a configuration space generates a real valued time series $\{f(x_1), f(x_2), \ldots\}$ whose autocorrelation function is given by

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

with variance $\sigma^2 = \langle f^2 \rangle - \langle f \rangle^2$. A landscape autocorrelation function, however, should yield information about the average changes of $f(x_{t+s})$ as the configuration space distance d of x_{t+s} to some reference point x_t is varied. This leads to the definition:

$$\rho(d) = \frac{\langle f(x) f(y) \rangle_{d(x,y)=d} - \langle f \rangle^2}{\sigma^2}, \tag{2}$$

where $\langle\, . \,\rangle_{d(x,y)=d}$ denotes an average over all pairs of configurations (x, y) at distance d from each other.

The autocorrelation of the landscape and the autocorrelation along a walk on the landscape are related to each other via

$$r(s) = \sum_d \varphi_{sd} \rho(d) \tag{3}$$

where φ_{sd} denotes the probability that a walk of s steps ends at a distance d from the starting point. Equation (3) establishes a complete correspondence between the random walk and the landscape autocorrelations, provided the φ_{sd} are independent of the initial conditions of the walk. For this to be the case a sufficient regularity of the configuration space is required. Such a regularity is missing, for example, on spaces of sequences with variable length that result from insertion and deletion operations. On sequence spaces with fixed length we obtain [4]

$$\varphi_{sd} = \varphi_{s-1,d-1} \frac{\nu - d + 1}{\nu} + \varphi_{s-1,d} \frac{d}{\nu} \frac{\kappa - 2}{\kappa - 1} + \varphi_{s-1,d+1} \frac{d+1}{\nu} \frac{1}{\kappa - 1} \tag{4}$$

In the generic case of CMs, where the range of f could be a non-scalar object - for example a structure - we must generalize the definition of $\rho(d)$ with the help of the distance measure on $\mathcal{Y}$. We propose the following

Definition.

$$\rho(d) = 1 - \frac{\langle d_{\mathcal{Y}}[f(x), f(y)]^2 \rangle_{d(x,y)=d}}{\langle d^2_{\mathcal{Y}} \rangle}. \tag{5}$$

It is readily seen that for landscapes, $d_{\mathcal{Y}}(a,b) = |a-b|$, we recover the usual autocorrelation:

$$\begin{aligned}
&\langle d^2_{\mathcal{Y}} \rangle - \langle d_{\mathcal{Y}}[f(x), f(y)]^2 \rangle_{d(x,y)=d} = \\
&\quad = \langle (f(p) - f(q))^2 \rangle - \langle (f(x) - f(y))^2 \rangle_{d(x,y)=d} = \\
&\quad = 2\langle f^2 \rangle - 2\langle f \rangle^2 - 2\langle f^2 \rangle + 2\langle f(x) f(y) \rangle_{d(x,y)=d} = \\
&\quad = 2\langle f(x) f(y) \rangle_{d(x,y)=d} - 2\langle f \rangle^2
\end{aligned} \tag{6}$$

Here $\langle (f(p) - f(q))^2 \rangle = 2(\langle f^2 \rangle - \langle f \rangle^2)$ denotes the average over pairs of randomly chosen configurations p, q.

Further insight into the statistical properties of CMs is provided by density surfaces [5]. A density surface $P(t|s)$ is the conditional probability that the images $f(x)$ and $f(y)$ have distance $d_{\mathcal{Y}}[f(x), f(y)] = t$, given that the configurations x and y are at distance $d_{\mathcal{X}}[x,y] = s$ from each other. The density surface shows how, and how fast, the distribution of image differences changes as the configurations become uncorrelated. This approach provides more information than can be obtained by random walks alone. The autocorrelation $\rho(d)$, equation (5), is extracted from the density surface with the following correspondences:

$$\begin{aligned}
\langle d_{\mathcal{Y}}[f(x), f(y)]^2 \rangle_{d(x,y)=s} &= \sum_t t^2 P(t, | \, s) \\
\langle d^2_{\mathcal{Y}} \rangle &= \sum_t \sum_d t^2 P(t|s) p(s).
\end{aligned} \tag{7}$$

$p(s)$ is the frequency of configuration pairs with distance $d_{\mathcal{X}} = s$. For sequence spaces this amounts to

$$p(s) = \frac{1}{\kappa^n} \cdot (\kappa - 1)^s \binom{n}{s}. \tag{8}$$

From the density surface the number of neighbors with identical image is retrieved as $P(0|1)(\kappa - 1)n$.

III Classification by Autocorrelation

It has been suggested [6] to characterize CMs by the behavior of the autocorrelation function at small distances $d_{\mathcal{X}}$. Landscapes with known autocorrelation function are listed in Tab.1.

Let the scaled distance be $\xi = d_{\mathcal{X}}/\max(d_{\mathcal{X}})$. Many combinatorial optimization problems exhibit landscape autocorrelation functions that can be approximated for small ξ by the first terms in the expansion of the exponential,

$$\rho(\xi) = 1 - \frac{1}{\alpha} \cdot \xi + \dots \tag{9}$$

The examples of table 1 all induce random walks with exponential autocorrelation functions $r(s)$. This need not be generally the case, as exemplified by superpositions of independent landscapes with different characteristic lengths that obviously result in landscapes with no unique α. Even if a walk on a landscape does not exhibit exactly a decaying exponential as autocorrelation function, a rough correlation indicator is given by the nearest-neighbor correlation:

$$\lambda = n\alpha \approx \frac{\langle d^2_{\mathcal{Y}} \rangle}{\langle d_{\mathcal{Y}}[f(x), f(y)]^2 \rangle_{d(x,y)=1}} \tag{10}$$

Another class of landscapes is characterized by discontinous autocorrelation functions, as for example, the random energy model [7], or the asymmetric TSP [8], see table 1.

Smooth landscapes have $\rho(\xi) = 1 - \xi^2 + \dots$, and nontrivial fractal landscapes are characterized by $\rho(\xi) = 1 - |\xi|^\alpha + \dots$, with $\alpha \neq 0, 1, 2$.

Table 1: Combinatorial optimization problems and their autocorrelation functions.

Name	Metric	$\rho(d)$	$r(s)$	ℓ	diamΓ	Ref.
REM	any	$\delta_{0,d}$	$\delta_{0,s}$	0	n	[7]
s.TSP	Tr.	?	$\sim e^{-4s/n}$	$n/4$	$n-1$	[8]
	2opt	?	$\sim e^{-2s/n}$	$n/2$	$n/2 \ldots n-1$	[8]
	c.Tr.	?	$\sim e^{-2s/n}$	$n/2$	$\frac{n(n-1)}{2}$	[6]
a.TSP	Tr.	?	$\sim e^{-4s/n}$	$n/4$	$n-1$	[8]
	2opt	?	$\sim \frac{1}{2}(\delta_{0,s}+e^{-2s/n})$	–	–	[8]
	c.Tr.	?	$\sim e^{-3s/n}$	$n/3$	$\frac{n(n-1)}{2}$	[6]
GM	Tr.	?	$\sim e^{-4s/n}$	$n/4$	$n-1$	[9]
GBP	Ex.	$1-\frac{n-1}{n-2}[8\frac{d}{n}-16(\frac{d}{n})^2]$	$(1-\frac{8}{n}+\frac{8}{n^2})^s$	$(n-3)/8$	$n/2$	[10]
LAS	Ham.	?	$\sim e^{-10s/n}$	$n/10$	n	[9]
rnd.Nk	Ham.	$(1-\frac{d}{n})(1-\frac{k}{n-1})^d$	$\sim(1-\frac{k+1}{n})^s$	$n/(k+1)$	n	[11,†]
p.r.Nk	Ham.	$(1-\frac{k+1}{n})^d$	$\sim(1-\frac{k+1}{n})^s$	$n/(k+1)$	n	[†]
adj.Nk	Ham.	equ.(24c)	$\sim(1-\frac{k+1}{n})^s$	$n/(k+1)$	n	[11,†]
p-Spin	Ham.	$1-\frac{2}{\binom{n}{p}}\sum'\binom{d}{j}\binom{n-d}{p-j}$	$\sim e^{-2ps/n}$	$n/(2p)$	n	[6]
SK	Ham.	$1-\frac{n}{n-1}[4\frac{d}{n}-4(\frac{d}{n})^2]$	$(1-\frac{4}{n})^s$	$n/4$	n	[6]

†) This paper.

Remark. $\sum'$ for the p-spin model denotes the sum over all odd j subject to the restriction $j{>}\min(d,p)$.
REM is Derrida's random energy model; s.TSP and a.TSP denote symmetric and asymmetric travelling salesman problems [12], GM is the graph matching problem [13]. (Mézard and Parisi 1986 The corresponding metrics are transpositions (Tr.), 2opt moves and canonical transpositions (c.Tr). GBP is the graph bipartitioning problem [14], its metric (Exc) is derived from exchanging a pair of objects. LAS stands for the low autocorrelated string problem [15]. The Sherrington-Kirkpatrick spinglass [16] is the special case $p{=}2$ of the p-Spin model introduced in [17] as a model for a rugged landscape in evolutionary optimization. The abbreviations rnd.Nk, p.r.Nk, and adj.Nk refer to random neighbour, purely random and adjecent neighbour N-k-model, resp. Here the canonical metric is the Hamming metric.

IV RNA Landscapes

IV.1 RNA Folding

An RNA sequence of length n is represented as a string $I = [s_1 s_2 \ldots s_n]$, with the s_i taken from an alphabet $\mathcal{A}$. Here we consider the natural four letter alphabet $\mathcal{A} = \{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$, the binary alphabets $\mathcal{A} = \{\mathbf{G}, \mathbf{C}\}$ and $\mathcal{A} = \{\mathbf{A}, \mathbf{U}\}$, as well as artificial four and six letter alphabets, $\mathcal{A} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ and $\mathcal{A} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{F}\}$, respectively, where complementarity is assumed between **A** and **B**, **C** and **D**, **E** and **F** (with energy parameters as in the **GC** case).

RNA structure can be broken down conceptually into a secondary structure, and a tertiary structure. The secondary structure is a pattern of complementary base pairings (figure 1). The tertiary structure is the three-dimensional configuration of the molecule. As opposed to the protein case, the secondary structure of RNA sequences is well defined; it provides the major set of distance constraints that guide the formation of tertiary structure, and covers the dominant energy contribution to the 3D structure. In this paper we will be concerned only with secondary structures.

The definition of secondary structure used in most computational approaches assumes planarity. Planarity means that unpaired nucleotides inside a loop cannot pair with unpaired nucleotides outside a loop. It is known, however, that unpaired bases from different loop regions may pair with each other, forming so-called pseudo-knots. While the computational problem for strictly planar secondary structures has been essentially solved in the early 80s [18 – 21], the problem involving long-range pseudo-knots is still unsolved.

A secondary structure $S(I) \in \mathcal{S}$ is formally defined as the set of all base pairs (s_i, s_j) with $(i < j)$ fulfilling two requirements [19]:

1) each base is involved in at most one base pair,

2) there are no knots or pseudo-knots, i.e. if (s_i, s_j) and (s_k, s_l) are base pairs then $i < k < l < j$ or $k < i < j < l$.

The basic elements of secondary structures are shown in figure 1. For the free energies of these building blocks experimental data are available. The elements are assumed to contribute additively to the overall free energy of the complete secondary structure. The data set used here has been taken from the literature [22 – 24]

Under thermodynamically reasonable assumptions the folding problem for planar RNA secondary structure consists in finding a minimum free energy structure, and can be attacked by the technique of dynamic programming. Our implementation of the folding algorithm follows the reasoning given by Zuker and Stiegler [20]. The combinatory map from the space of sequences, $\mathcal{A}^n$, into the space of minimum free energy structures, $\mathcal{S}$,

$$S : \mathcal{A}^n \to \mathcal{S}, I \to S(I), \tag{11}$$

is the quintuple $(\mathcal{A}^n, d_H, \mathcal{S}, d_{\mathcal{S}}, S)$ with d_H denoting the Hamming distance, and $d_{\mathcal{S}}$ being a suitably defined distance between secondary structures (see below). If we view the folding procedure as a map from a sequence to the free energy of the corresponding structure,

$$\Delta G : \mathcal{A}^n \to \mathrm{I\!R}, I \to \Delta G(I) \tag{12}$$

we obtain the combinatory landscape $(\mathcal{A}^n, d_H, \mathrm{I\!R}, |\,.\,|, \Delta G)$.

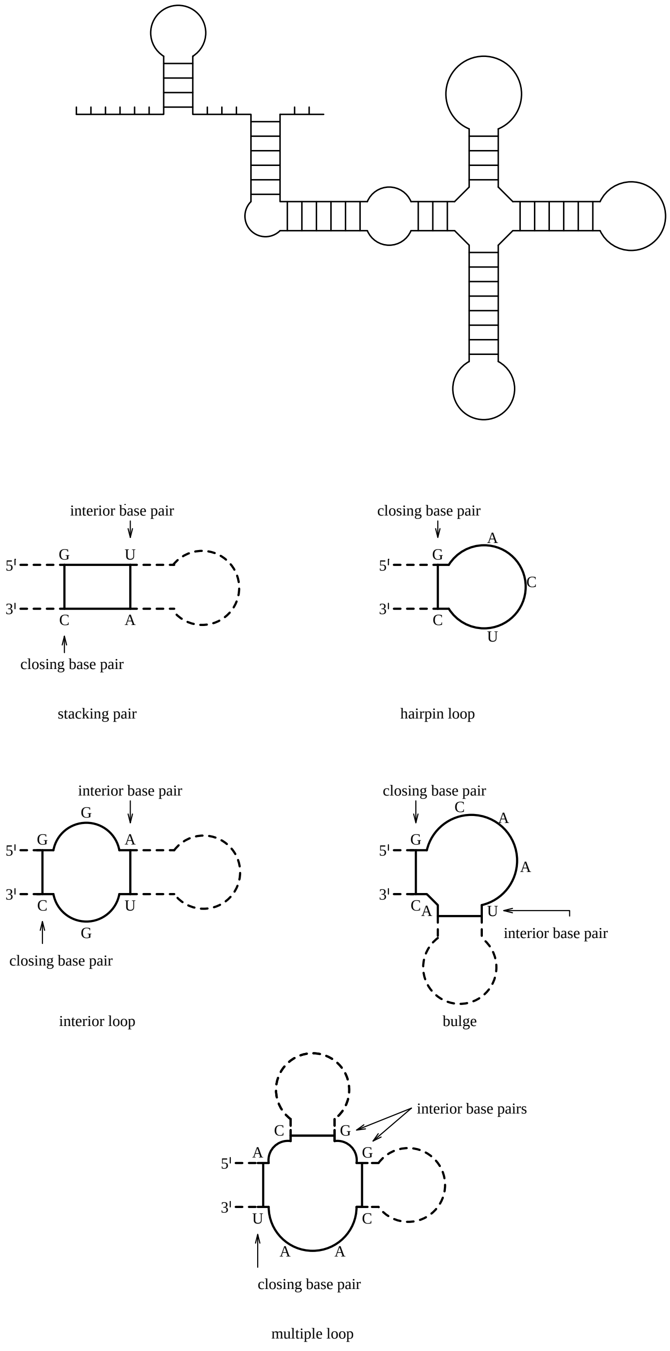

Figure 1: Top: Example of a secondary structure. Below: Basic structure elements. Every secondary structure can be decomposed into such basic elements. The free energy of a secondary structure is the sum over all contributions of these elements. Their energy contributions have been experimentally determined as a function of the nucleotide sequence [22–24].

IV.2 Distribution of Energies

Here we focus on global features relating to the distribution of free energy values over the RNA landscape. Table 2 shows the mean free energy values $\langle \Delta G \rangle$ and variances σ^2 for a variety of alphabets. We have also computed the skewness (third moment scaled by σ^3) and the kurtosis (fourth moment scaled by σ^4) of the distribution (not shown).

The main observations are that $\langle \Delta G \rangle$ scales linearly with chain length n for all alphabets, and so does the variance σ^2 for $n > 50$ (see table 4 below). The distribution sharpens considerably for longer chains. A Gaussian distribution is characterized by vanishing skewness and a value of 3 for the kurtosis. This seems to be mostly the case for the biophysical **GCAU** alphabet and long chains. The other alphabets show deviations from the Gaussian case in skewness and/or kurtosis. The available data do not allow a definite statement about the limiting behavior for very large n to be made. We did not pursue this issue in depth because of doubts about the validity of a thermodynamic (rather than kinetic) folding algorithm in that limit.

IV.3 Free Energy Autocorrelation

Figure 4a shows an example of the landscape autocorrelation function $\rho(d)$. Complementary sequences have similar structures and energies. For each reference sequence on binary alphabets there is only one complement, and, therefore, this approximate symmetry shows up as a U-shaped correlation. Table 3 lists the characteristic lengths of the landscape, equation (10), for various alphabets as a function of n. The characteristic length has been calculated by both random walks and density surfaces. The agreement is remarkable, indicating that the landscapes are indeed sufficiently regular for the random walk approach to be meaningful.

Table 2: Averages of free energies, ΔG, and their variances, σ^2, for 1) **GC**, 2) **AU**, 3) **GCAU**, 4) **GCXK** and 5) **ABCDEF**. For 3) the biophysical parameter set is used, for 4) and 5) the **GC** parameter set is used for all base pairs. Energy unit: 0.1 kcal·mol^{-1} Energy data set taken from [23].

	$-\Delta G$					σ				
n	1)	2)	3)	4)	5)	1)	2)	3)	4)	5)
20	74.40	2.22	10.29	19.06	6.59	28.80	5.12	14.63	19.95	12.35
25	108.85	5.68	17.56	31.50	16.05	31.86	8.78	18.56	23.88	16.05
28				39.38					25.79	
30	144.23	9.59	25.58	44.46	16.91	34.55	11.24	21.81	26.96	18.89
*30			25.68					22.00		
35	180.27	14.09	33.61	58.24	22.71	36.96	13.18	24.35	29.22	21.32
40	217.00	18.83	42.33	72.60	29.30	39.17	14.64	26.88	31.54	23.59
45	253.73	23.53	50.58	86.98	35.67	40.87	15.81	28.83	33.25	25.38
50	290.61	28.54	60.12	101.89	42.36	42.66	16.89	31.03	35.32	27.24
*50			60.16					31.13		
55	327.59	33.54	68.84	116.47	48.89	44.23	17.81	32.82	36.89	28.70
60	365.50	38.61	77.48	131.28	55.81	45.51	18.79	34.63	38.19	30.10
70	440.73	48.46	95.64	161.63	69.59	47.95	20.25	37.64	40.99	32.66
80	516.83	59.32	113.77	192.19	83.57	50.76	21.78	40.46	43.37	34.87
100	669.35	80.12	150.78	254.56	111.46	54.21	24.17	45.96	48.04	38.90
120	824.41					58.42				
150	1054.77	133.45	244.57			62.12	28.79	57.06		
200	1445.62	188.21	339.16			68.63	32.58	66.28		
250	1838.36	243.62	434.93			74.08	36.04	73.79		
300	2233.63	299.62	531.34			79.80	39.70	81.14		
350	2629.50	355.88	628.59			84.40	42.72	87.96		
400	3026.54	412.50	726.02			88.44	45.50	94.41		
450	3423.91	469.30	823.52			92.98	47.70	99.92		
500	3822.34	526.24	921.46			95.74	49.69	105.2		

*) Specially stable tetraloops taken into account.

Table 3: Correlation lengths for various types of RNA free energy landscapes. R and D denote calculations via random walks and with density surfaces via equation (7), respectively. Energy parameters are taken from [23].

n	GC		AU		GCAU		GCXK		ABCDEF	
	R	D	R	D	R	D	R	D	R	D
20	1.96	1.97		1.13	3.64	3.62	2.62	2.48	2.70	2.50
				1.27						
25	2.44	2.41		1.50	4.83	4.76	3.21	3.17	3.46	3.16
				1.44						
28							3.49			
30	2.95	2.96	1.85	1.93	5.97	6.01	3.84	3.75	4.02	3.92
		2.90			*6.08					
		2.82								
35	3.49	3.23		2.23	7.29	7.79	4.35	4.39	4.59	4.66
40	4.00	3.97	2.49	2.56	8.67	8.16	4.88	4.85	5.46	
		4.04				8.12				
45	4.47	4.15		2.75	10.05	9.60	5.38	5.22	5.96	5.97
50	4.57	4.51		3.07	11.49	11.11	6.08	5.95	6.68	6.32
				3.15	10.83		5.61			
					*11.74					
55		5.43		3.24	12.10		6.68			7.05
60	5.50	6.17		3.72	14.88	13.20	7.05	7.03	7.97	7.55
70		6.68		4.45	16.22	16.40		7.68	9.06	8.71
		6.56		4.19		15.05		8.09		
		7.03								
80		7.82		4.88	19.03			8.84	9.98	9.89
				4.89						
90				5.14					11.45	
100	8.41	8.52		6.11	24.73	25.18	12.20	10.57	12.31	11.55
120		10.55								

$^{\star)}$ With specially stable tetra-loops.

The characteristic length scales linearly with n. This is understood by observing that the absolute free energy difference between two neighbors (one-digit mutants) is bounded from above by twice the largest possible stacking energy plus the strain of an internal loop of size two (approximately $10\,\text{kcal}\cdot\text{mol}^{-1}$). This is seen by considering that in the worst case a mutation disables a base pair, thus interrupting a stack by creating an internal loop. If this leads to a complete refold of the sequence, then the new structure must have a smaller absolute energy difference than the worst case. The expectation value of the squared energy difference between nearest neighbors will, therefore, level off to a constant at sufficiently large n. At the same time the data of the previous section show a variance linear in n, and hence we have $\lambda \sim n$ (see table 4 below).

Figure 2 shows the free energy density surface

$$P(z|d) = P(|\Delta G(I_1) - \Delta G(I_2)| = z \,\big|\, d_H(I_1, I_2) = d). \tag{13}$$

It reveals a very rugged landscape. Cuts along constant Hamming distance show roughly the positive half of a Gaussian, except for a pronounced peak at $P(0|1)$ corresponding to neighbors with identical energy. If the free energies were sums of independently distributed random variables one would expect for the density surface:

$$P(z|d) = \frac{1}{\sqrt{\pi(1-\rho(d))}\,\sigma} \cdot \left\{ \exp\left(-\frac{z^2}{4\sigma^2(1-\rho(d))} \right) - \delta_{0,z} \right\} \tag{14}$$

Figure 2b shows $P(z|d)$ as in equation (14) with $\rho(d)$ taken from the data in table 3. The overall shape is in agreement. For comparison we also plot the corresponding exact density surface for the Sherrington-Krikpatrick spin glass. ($\rho(d)$ from table 1, and references therein.)

From the density surface we can read off the frequency of neutral neighbors (see section 2). This is shown in figure 3. The fraction of neutral neighbors tends to a positive constant for long chains. From statistics of structural

Table 4: Scaling with sequence length of average free energy, variance and correlation length of free energies for different alphabets. Energy parameters from [23].

	GC	**AU**	**GCAU**	**GCXK**	**ABCDEF**
			$-\langle \Delta G \rangle$		
slope	7.8120	1.0948	1.9006	2.9495	1.3044
intercp	−101.21	−26.14	−32.25	−43.97	−21.40
corr.	0.99996	0.9998	0.99992	0.9996	0.998
			σ^2		
slope	17.33	5.06	22.61	23.56	17.07
intercp	959.8	22.2	−180.4	16.1	−143.3
corr.	0.996	0.997	0.99996	0.997	0.998
			λ		
slope	0.0857	0.0600	0.2627	0.1078	0.1182
intercp	0.43	0.06	−1.93	0.50	0.47
corr.	0.992	0.997	0.997	0.993	0.997

elements [5,25] we know that for long chains the average loop size becomes constant, while the average number of loops per structure scales linearly with n. The sequence positions that can be changed without affecting the structure and its energy must be located within loops (or external elements). This explains the linear scaling of the number of neutral neighbors.

IV.4 Structure Autocorrelation

As detailed in section 2, computing the autocorrelation of the structures themselves requires a distance measure on the space of structures. The

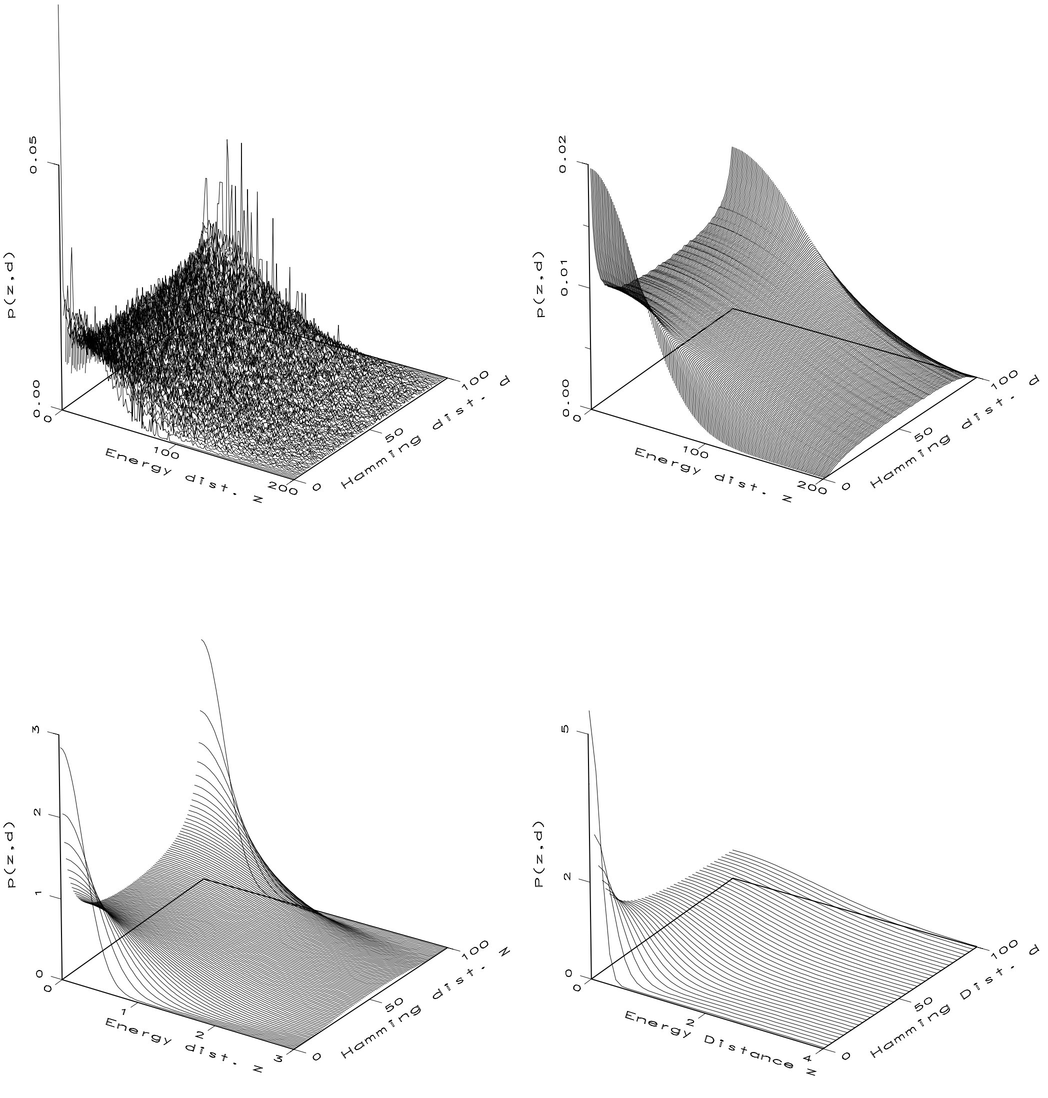

Figure 2: **a)** Sampled denisty surface $P(z|d)$ for energy differences of minimum free energy structures on **GC**-sequences of length n=100. **b)** Gaussian approximation with the statistically determined autocorrelation function $\rho(d)$). **c)** Exact density surface for the Sherrington Kirkpatrick spin-glass. **d)** Exact density surface for a RN n-k model with k=7.

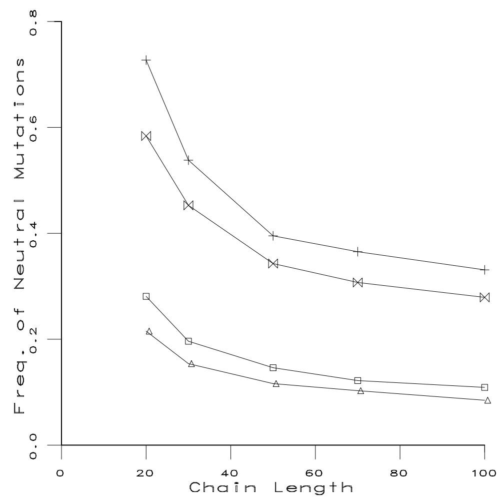

Figure 3: Frequency of neutral neighbors $(0|1)$ on the free energy landscape as a function of chain length. The upper two of curves belongs to the **GCAU** alphabet, the lower two curves belong to the **GC** alphabet. The upper curve in each pair refers to the older parameter set taken from [22], the lower curve in each pair refers to the updated set taken from [23].

definition is essentially based on converting one-to-one secondary structure graphs into rooted ordered trees [5,26 – 28] The distance between two trees is obtained as a generalization of sequence alignment procedures, and involves minimizing the cost of transforming one tree into the other with elementary editing operations, like deletion, insertion, and relabeling of nodes. For details see [5].

Fig. 4b shows the structure autocorrelation function $\rho(d)$. The U-shaped form for the binary alphabet arises for the same reason given in sect. IV.3.

The structure density surfaces of sequences with $n = 100$ for the **GC** and **GCAU** alphabet are shown in figure 5. First we note that nearest neighbors in configuration space can exhibit already substantially different structures with significant probability. Furthermore, it is extremely unlikely that two randomly chosen sequences fold into identical structures. This is in

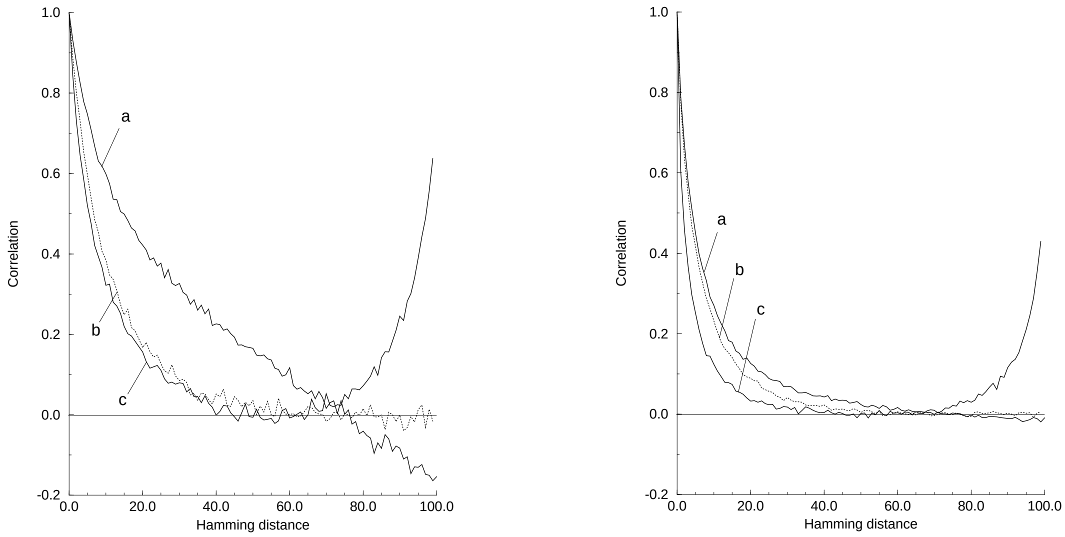

Figure 4: Autocorrelation functions for sequences of length n=100 for free energy landscapes (left) and (right) secondary structure landscape with alphabets **GCAU** (**a**), **GCXK** (**b**), and **GC** (**c**).

sharp contrast to the energy landscape. Stated differently: there are many more structures than energies. Nonetheless, the distribution of structure distances approaches already at fairly short Hamming distances the distribution expected for random sequences. In contrast to the free energy there is no size independent upper bound to the distance between secondary structures of two neighboring sequences. This suggests that in a small ball around a random sequence one may find the vast majority of all minimum free energy structures $\mathcal{S}$ [5]. The phenomenon is known as "shape space covering" [29] We will report elsewhere on the verification of this conjecture with additional methods. Although there are many fewer different energy values than minimum energy structures, the energies are not scattered randomly across the sequence space.

The characteristic lengths of the structure autocorrelation (table 5) are shorter by a factor between 2 and 3 compared to their free energy counterparts. They roughly correspond to the Hamming distance at which the

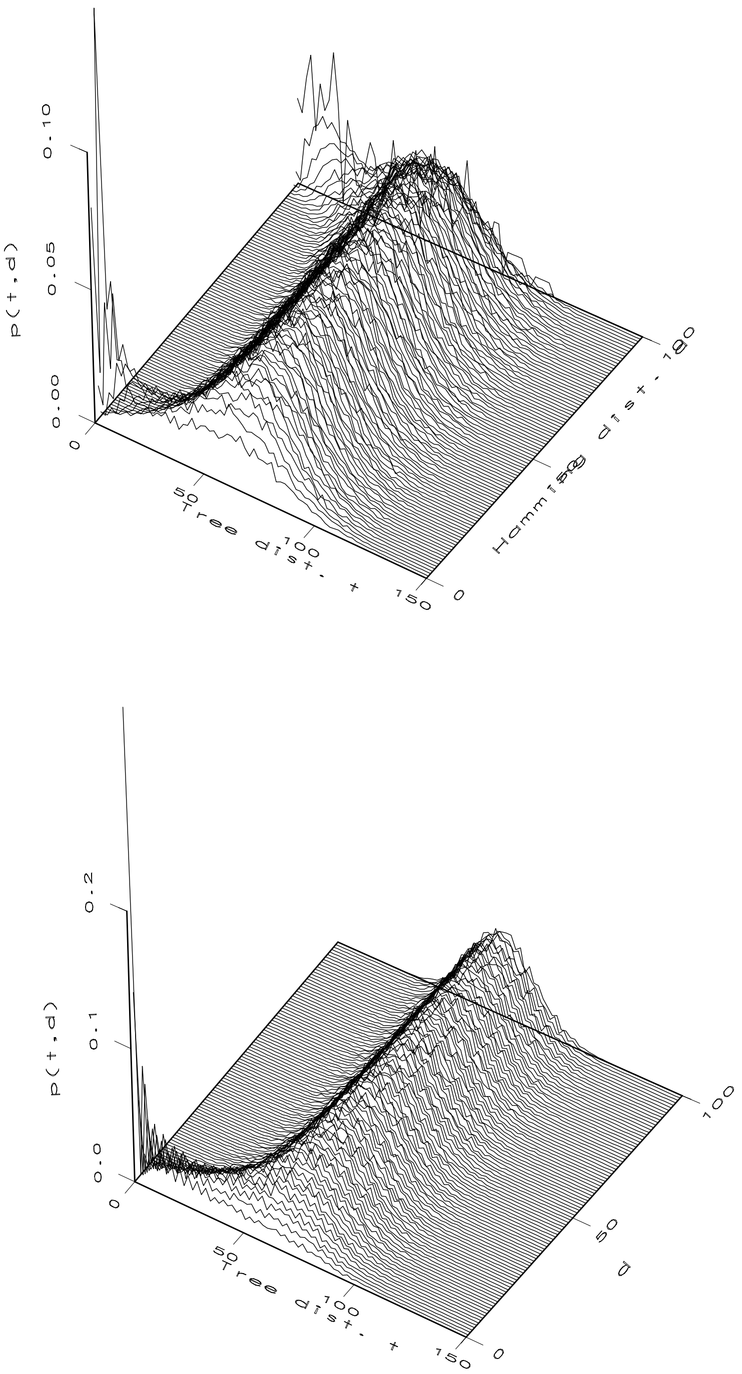

Figure 5: Density surface for secondary structure differences as determined by tree editing. Parameters are as in figure 2. n=100, upper part: **GC**, lower part: **GCAU**

Table 5: Characteristic lengths of RNA secondary structure landscapes for various alphabets and chain lengths. Energy parameters are taken from [23].

n	**GC**	**AU**	**GCAU**	**GCXK**	**ABCDEF**
20	1.51	1.22	2.84	2.43	2.50
25	1.69	1.61	3.68	2.88	3.28
30	1.87	1.85	3.99	3.25	3.92
35	2.02	2.12	4.49	3.57	4.51
40	2.15	2.24	4.85	3.74	
45	2.33	2.33	4.97	4.10	5.49
50	2.35	2.50	5.46	4.33	5.84
55	2.62	2.64			6.20
60	2.68	2.83	6.01	4.84	6.60
70	2.87	2.91	6.25	5.24	7.14
80	3.13	3.32		5.67	7.93
90		3.60			
100	3.36	3.77	7.63	6.45	8.94
120	3.72				

dominant peak resulting from identical or very similar structures has dissolved. Randomization of structure for **GCAU** sequences of length $n = 100$ occurs already at a distance of about $d_H = 16$ from a random reference point (figure 5). From table 5 we see that this corresponds to a ball of radius $d_H \approx 2\lambda$. The characteristic length data are roughly consistent with linear scaling in n.

The ordering of characteristic lengths of the structure autocorrelation for long sequences is summarized by:

$$\mathbf{AU} \approx \mathbf{GC} \ll \mathbf{GCXK} < \mathbf{GCAU} < \mathbf{ABCDEF} \tag{15}$$

Binary alphabets generally form more structures because the probability for two randomly chosen position along the sequence to pair is highest. Changing one position, therefore, is more likely to alter the minimum energy structure. In contrast, the characteristic lengths of the free energy autocorrelation does not follow the same ordering:

$$\mathbf{AU} < \mathbf{GC} < \mathbf{GCXK} < \mathbf{ABCDEF} \ll \mathbf{GCAU} \tag{16}$$

IV.5 Biased Walks and Local Optima

Additional information about the local structure of landscapes is provided by biased walks. These are random walks aimed at reaching local extrema of the landscape. The data provided by this technique will be compared with a statistical model in the next sections.

We define a local minimum $\hat{y}$ in configuration space by

$$f(\hat{y}) \leq f(x) \qquad \text{for all neighbors } x \text{ of } \hat{y}. \tag{17}$$

Local maxima are defined analogously. The term local optimum means either minimum or maximum according to context.

We consider here two types of biased walks, adaptive walks and gradient walks. In an adaptive walk a starting point x_0 is chosen at random. The walk then proceeds to a randomly chosen neighbor x' such that $f(x') < f(x_0)$, and it stops if no neighbor satisfies this condition, at which point the walk has reached a local minimum. A gradient walk, in contrast, proceeds to the neighbor x' for which $f(x')$ is minimal.

Table 6 compiles the average lengths of adaptive and gradient walks on the free energy landscape, as well the the average free energy (and the corresponding variances) of start and end points. These quantities were calculated for an energy parameter set available in the late 70s [22] and for a recently updated parameter set [23,24]. While the absolute values differ, the scaling behavior is not affected. The average free energy of local

Table 6: Adaptive and gradient walks on RNA free energy landscapes. $\bar{F}$ denotes the free energy of random sequences. Average free energies of local optima are denoted by $\bar{F}_{\mathrm{opt}}$, $\bar{F}^{\star}_{\mathrm{opt}}$, and $\bar{F}^{\star\star}_{\mathrm{opt}}$, depending on whether they are generated randomly, as end points of adaptive walks, or as end points of gradient walks, respectively. The standard deviations of their distributions are denoted by σ_F, σ_{opt}, $\sigma^{\star}_{\mathrm{opt}}$, and $\sigma^{\star\star}_{\mathrm{opt}}$. The lengths of adaptive and gradient walks are denoted by L_{adap} and L_{grad}, respectively. The rightmost column compiles the probability $p_{\mathrm{l.o.}}$ for finding a local optimum at random.

n	$\bar{F}$	σ_F	$\bar{F}_{\mathrm{opt}}$	σ_{opt}	$\bar{F}^{\star}_{\mathrm{opt}}$	$\sigma^{\star}_{\mathrm{opt}}$	$\bar{F}^{\star\star}_{\mathrm{opt}}$	$\sigma^{\star\star}_{\mathrm{opt}}$	λ	L_{adap}	L_{grad}	$p_{\mathrm{l.o.}}$
GC												
Energy parameters from [22]												
20	118.9	43.0	179.7	32.0	204.5	29.1	211.5	31.2	2.10	2.42	1.72	0.1138
25	178.9	47.3	257.6	44.1	300.4	37.8	305.2	42.0	2.70	3.30	2.29	0.0530
30	239.7	50.8	329.3	50.6	398.5	48.7	402.9	53.7	3.09	4.44	3.12	0.0220
35	303.1	54.0	390.3	43.6	493.1	60.5	496.5	67.4	3.57	5.49	3.88	0.0108
40	368.6	57.6	465.5	48.8	584.7	74.5	585.3	79.1	3.98	6.42	4.48	0.0084
50	500.1	62.3	633.4	48.9	759.8	90.7	755.5	92.4	4.67	8.06	5.50	0.0025
60	632.2	66.7	768.0	–	934.4	97.4	930.9	94.9	5.65	9.61	6.42	0.0008
Energy parameters from [23]												
20	74.4	28.8	127.9	24.9	148.3	21.6	152.0	22.2	1.96	3.55	2.51	0.0230
30	144.2	34.6	213.1	28.1	262.6	36.6	265.1	37.3	2.93	5.32	3.70	0.0050
40	217.0	39.2	315.0	29.5	369.2	44.4	372.4	44.7	4.00	6.78	4.69	0.0011
50	290.6	42.7	415.0	–	478.0	48.9	481.9	49.3	4.51	8.27	5.68	0.0002
60	365.5	45.5	–	–	587.6	56.6	589.4	55.1	6.17	9.62	6.64	–
GCAU												
Energy parameters from [22]												
30	34.2	31.0	94.6	45.4	208.5	72.3	228.5	78.4	6.73	8.01	5.12	0.0009
40	59.8	37.9	–	–	307.3	93.5	339.8	103.7	9.85	11.62	7.57	–
50	90.6	43.8	–	–	404.1	105.1	441.8	112.0	13.76	14.77	9.31	–
60	128.1	48.1	–	–	504.6	118.0	–	–	18.38	17.99	–	–
70	162.7	50.6	–	–	604.2	125.3	644.1	132.9	20.80	21.22	13.32	–
80	201.9	55.3	–	–	704.2	139.2	–	–	23.81	24.17	–	–
Energy parameters from [23]												
30	25.7	22.0	–	–	172.9	43.9	181.1	47.4	6.08	11.84	7.17	–
40	42.3	26.9	–	–	253.0	50.4	272.0	54.8	8.14	16.26	9.95	–
50	60.2	31.1	–	–	325.8	59.6	352.3	65.1	11.74	20.73	12.53	–
60	77.5	34.6	–	–	401.8	68.3	428.0	71.9	13.20	25.30	14.92	–
70	95.6	37.6	–	–	478.4	73.1	508.0	80.1	16.20	29.13	17.44	–

optima reached by adaptive and gradient walks is roughly the same and scales linearly with n. The rightmost column in table 6 summarizes the frequency $p_{\mathrm{l.o.}}$ with which the initially chosen sequence was already a local optimum. We refer to such a sequence as a "random optimum". The frequency $p_{\mathrm{l.o.}}$ decreases exponentially with chain lenght n (see sect. IV.5). This frequency was, however, too low to yield meaningful quantities for **GCAU** sequences in both parameter sets. We note that the energy of local optima reached at the end of the biased walks is substantially lower than the average free energy of random optima.

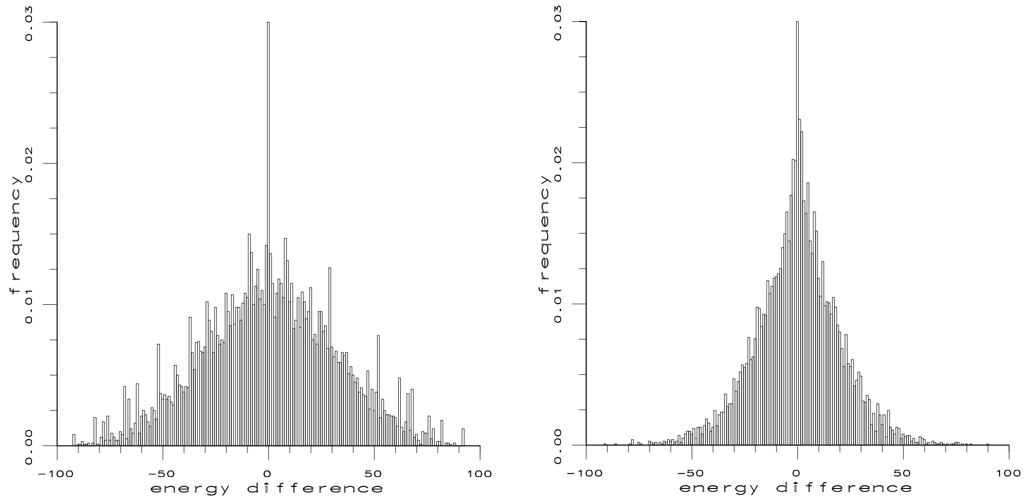

Figure 6: Distribution of per-step energy differences along a random walk. (a) **GC** sequences of length n=100. (b) **GCAU** sequences of length n=100.

The distribution of free energy increments per step along a random walk is shown in Figure 6. The local deviations from a Gaussian distribution are again most pronounced at 0, resulting from the substantial fraction of neutral neighbors. In the case of binary sequences (figure 6a) particular free energy changes are favored. As a biased walk proceeds the distribution of energy increments is obviously altered. After some five steps the distribution is dominated by a few quanta (figure 7) that identify the most likely structural changes.

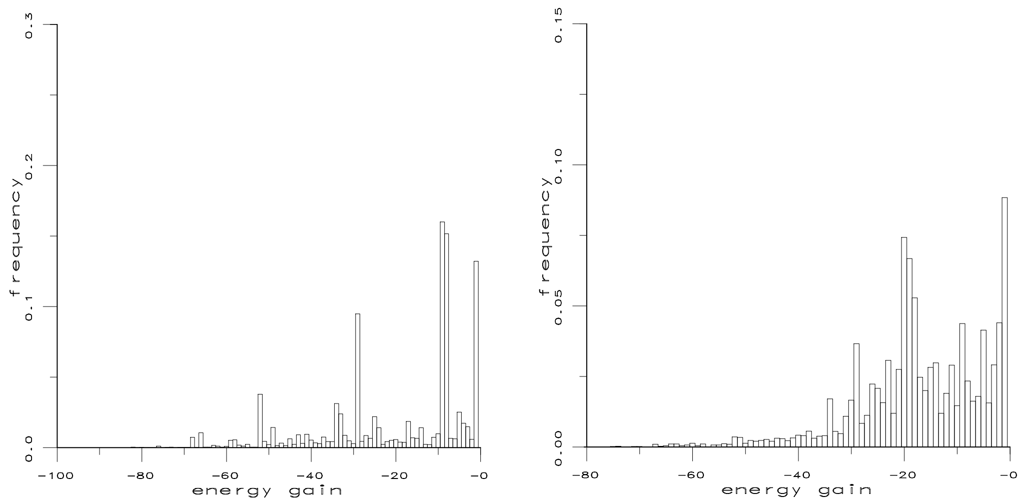

Figure 7: Distribution of energy increments after more than 4 steps along a gradient walk. **(a) GC**-landscape and **(b) GCAU**-landscape.

In figure 8 we plot the average change in free energy for landscapes of different length along biased walks. The number of walk steps is scaled by the corresponding characteristic length λ. The figure demonstrates that for a fixed alphabet the properly scaled statistical features of walks are independent of the system size.

V Kauffman's n-k Model

Because of the important role of landscapes in problems of optimization and evolutionary adaptation [30] there has been considerable interest in devising simple statistical models. The hope being that these models reflect

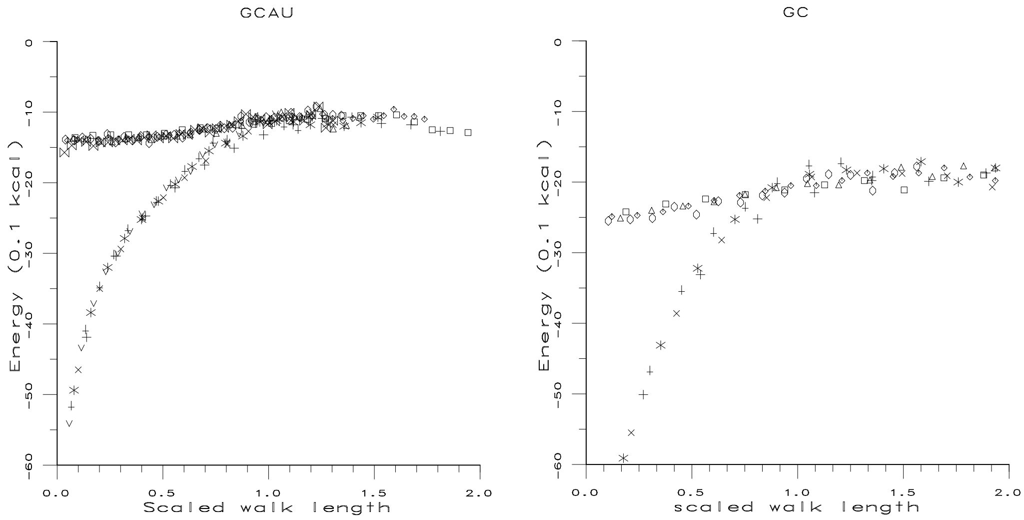

Figure 8: Average energy increment per step for biased walks on the **(a) GC**-landscape and **(b) GCAU**-landscape. $\square$, $\triangle, \diamond, ()$, $\bowtie$ belong to adaptive walks and chain lengths 30, 40, 50, 60, resp. The corresponding gradient walks are denoted by $+, \times, *, \dagger$ and $\vee$. Energy parameter set taken from [23].

the proper qualitative features of their natural counterparts. In the previous section we had a close look at an expensive but realistic model of a landscape induced by RNA folding. In this section we present a frequently used statistical model due to S. Kauffman, the so-called n-k model [31,32].

Let us assume that the energy F of a string of n bits is the average of contributions from each of the individual bits. We choose the contribution from the i^{th} bit, f_i, as a random function of the state of that bit and a context given by $k < n$ other bits. Each of the 2^{k+1} possible values for f_i, one for each of the possible states of the $k+1$ bits upon which f_i depends, is assigned by selecting an independent random variable from some specified probability distribution, $P(x) = \Pr(f_i \leq x)$. This set of assignments constitutes the "energy table" for the i^{th}bit. There is a different, independently generated table for each of the n bits, which, once chosen, is never reassigned.

It remains to specify the k sites which influence a given position. The simplest choice is to imagine that the neighborhoods consist of the $k/2$ bits that are the nearest neighbors on each side of the bit in question and to assume that the bits are arranged in a circle (periodic boundary conditions). The other extreme is to pick the k sites at random. In a variant of this model it is not required that f_i depends on i, but all $k+1$ relevant sites are chosen randomly. These three versions of the n-k model will be referred to as AN (adjacent neighborhood), RN (random neighborhood) and PR (purely random) model, respectively. These two extremes correspond to two important types of spin glasses: the adjacent neighborhoods correspond to a one dimensional, short range spin glass; the random neighborhoods correspond to a dilute, long range spin glass.

If the underlying probability distribution of the site energies, f_i, has a finite variance, the distribution of the fitness of the entire string,

$$F = \frac{1}{n}\sum_i f_i, \tag{18}$$

will tend to a Gaussian with mean μ and variance σ^2/n, where μ and σ^2 are, respectively, the mean and variance of the f_i's as $n \to \infty$.

V.1. Correlation

Weinberger [3] shows that the autocorrelation function of a random walk on the n-k model landscape is a single decaying exponential to within an error of $\mathcal{O}(1/n)$. All isotropic landscapes of this kind can be generated by an appropriate choice of the mean and variance of the site energy distribution, $P(x)$, and an appropriate choice of k.

The ruggedness of the landscape varies dramatically as k varies from 0 through $n-1$. For $k=0$, each site is independent of all other sites. The autocorrelation function in this special case reads

$$\rho(d) = 1 - d/n \tag{19}$$

After some algebra one shows that a random walk on this landscape generates an AR(1) process with autocorrelation function

$$r(s) = \sum_{d=0}^{s} \varphi(s,d)\rho(d) = \left(1 - \frac{1}{n}\frac{\kappa}{\kappa-1}\right)^{s}. \tag{20}$$

In contrast, the $k = n-1$ landscape is the random energy model: the energy contribution of each site then depends on all of the other sites because the context for each of the $n-1$ other bits is changed when even a single bit is flipped. In this case, therefore, the energy of each n bit string is assigned a energy that is statistically independent of its neighbors.

The autocorrelation functions for the various types of n-k models are obtained from

$$[F(x)-F(y)]^2 = \sum_{i=1}^{N}[f_i(x)-f_i(y)]^2 + 2\sum_{i<j}[f_i(x)-f_i(y)][f_j(x)-f_j(y)]. \tag{21}$$

The average over the second sum vanishes because of the independence of the f_i's and the contributions of the terms $(f_i(x)-f_i(y))^2$ are zero if x and y have the same digit at all positions which are in the neighborhood of site i. Otherwise the contributions are random with mean value σ^2. Thus we obtain

$$\langle (F(x)-F(y))^2 \rangle = \frac{1}{n}2\sigma^2 n h(d), \tag{22}$$

where $h(d)$ is the probability for a site to appear at least once in the list of the d sites which are changed by moving from x to y. For the n-k model it is evident that complementary strings are uncorrelated because they do not have any contribution to site fitnesses in common, therefore the variance of F is σ^2 and

$$\rho(d) = 1 - h(d). \tag{23}$$

For the three variants of the n-k model the functions $h(d)$ are derived in the appendix.

$$\begin{aligned} h_{RN}(d) &= 1-(1-\frac{d}{n})(1-\frac{k}{n-1})^d \\ h_{PR}(d) &= 1-(1-\frac{k+1}{n})^d \\ h_{AN}(d) &= \frac{k+1}{n}d - \frac{1}{\binom{n}{d}} \sum_{j=1}^{\min(k,n+1-d)} (k-j+1)\binom{n-j-1}{d-2} \text{ for } d \geq 2 \end{aligned} \tag{24}$$

Evidently, $h(d)$ is independent of the number of letters in the alphabet. For all three types of n-k models the correlation length can be estimated from the nearest-neighbor correlations, equation (10),

$$\rho(1) = (1-\frac{1}{n})(1-\frac{k}{n-1}) = 1-\frac{k+1}{n} \tag{25}$$

since all $h(1)$ are identical. This yields the estimate

$$\lambda = -1/\ln(1-\frac{k+1}{n}) = \frac{n}{k+1} - \frac{1}{2} + \mathcal{O}(\frac{k+1}{n}). \tag{26}$$

V.2. Biased Walks and Local Optima

In the case $k=0$ there is, with probability one, a unique optimal digit for each site; hence, a single specific sequence comprised of the optimal digit values in each position is almost surely the single, global optimum in the energy landscape. Any other string is suboptimal, and lies on a connected walk via nearest neighbor better variants to the global optimum. The length of the walk is just the Hamming distance from the initial string to the global optimum. For a randomly chosen initial string, $(\kappa-1)/\kappa$ of the digits will be in a energetically less favourable state (κ is the size of the alphabet), hence the expected walk length is just $\frac{\kappa-1}{\kappa}n$.

For the random energy model it has been shown [1,33 – 35] that the landscapes have many local optima, that the walks to optima are short ($\mathcal{O}(\ln n)$), and that only a small fraction of local optima are accessible from any initial string.

Weinberger [36] derived expressions for the expected number of local optima and the probability distribution of their energies. In the Gaussian case, the expected number of local optima, $\mathcal{N}_{\text{l.o.}}$, is given by

$$\log(\mathcal{N}_{\text{l.o.}}) \approx c_0 + n \cdot \left[\log \kappa - \frac{\log[(\kappa-1)(k+1)]}{k+1}\right] \tag{27}$$

where c_0 is some small constant. The distribution of energies of local optima is Gaussian with mean $\langle F_{l.o.}\rangle = \mu - \eta\sigma$, where

$$\eta = \sqrt{\frac{2\log[(\kappa-1)(k+1)]}{k+1}} \tag{28}$$

and variance

$$\sigma^2_{\text{l.o.}} = \sigma^2 \cdot \frac{1}{n} \frac{1}{1+4(\kappa-1)(k+2)\frac{\log(k+1)}{k+1}}. \tag{29}$$

Estimates for the average lengths of biased walks have been derived recently [36]. The average length of gradient walks can be estimated from the average distance of two local optima as

$$L_{\text{Grad}} = n \cdot \left(\frac{\kappa-1}{\kappa}\right) \frac{\log[(\kappa-1)(k+1)]}{(k+1)\log\kappa}. \tag{30}$$

As an estimate for the length of adaptive walks one obtains

$$L_{\text{Adap}} = n \cdot \frac{\log[(\kappa-1)(k+1)]}{(k+1)}. \tag{31}$$

The lengths of adaptive walks and gradient walks differ, therefore, by a constant factor

$$\frac{L_{\text{Adap}}}{L_{\text{Grad}}} = \frac{\kappa\log\kappa}{\kappa-1} \tag{32}$$

The chance to find a local optimum at random is given by

$$\log p_{\text{l.o.}} = \log\mathcal{N}_{\text{l.o.}} - n\log\kappa = c_0 - n \cdot \frac{\log[(\kappa-1)(k+1)]}{k+1}. \tag{33}$$

VI Comparison of the RNA Landscape with the n-k model

In order to compare the n-k model and the RNA landscape we need a hypothetical k-value. From equation (26) we obtain

$$k = n\left(1 - e^{-1/\lambda}\right) - 1. \tag{34}$$

Since the characteristic length of the RNA landscape scales linearly with n, $\lambda \approx \alpha \cdot n + a_0$, we find asymptotically

$$k = \frac{1}{\alpha} - 1 + \mathcal{O}(n^{-1}). \tag{35}$$

Table 7 compares the statistics of local optima in the n-k landscape with the statistics of local optima in the RNA free energy landscape. The ensembles of local optima in the RNA case have been obtained with gradient and adaptive walks as detailed in sect IV.5. The comparisons use the scaled quantities

$$f_{\text{l.o.}} = \frac{\langle F\rangle - \langle F_{\text{l.o.}}\rangle}{\sigma}. \tag{36}$$

Table 7 shows that average free energies of local optima for the RNA landscapes, sampled with both gradient ($f^{\star\star}_{\text{l.o.}}$) and adaptive walks ($f^{\star}_{\text{l.o.}}$), differ significantly from those predicted by the n-k model. Due to the sampling method the local optima in the RNA case are obviously not a true random sample, but are biased towards those optima that have large basins of attractions. These optima usually have more extreme values than purely randomly sampled local optima. The comparison, therefore, is only limited. On the other hand, the n-k model predicts a distribution of average energies of local optima that sharpens like $\sigma^2_{\text{l.o.}}/\sigma^2 \sim 1/n$, equation (29). If this were the case for the RNA landscape we should observe no difference between our biased sample and a random sample. This is clearly not the case. In fact, the variances $\sigma^2_{\text{l.o.}}$ (table 7), scale linearly with chain length. In the **GC** case we

Table 7: Hypothetical k-values and a comparison between the n-k model and the computed RNA data from table 6. Small numbers refer to the older data set from [22].
k_{pred} is the predicted valued of k for the RNA landscape as obtained from equation (34), f_{pred} is the predicted average scaled energy of a local optimum according to equation (28). $f_{\text{l.o.}}$, $f^{\star}_{\text{l.o.}}$, and $f^{\star\star}_{\text{l.o.}}$ are the numerical average scaled energies for random local optima and for local optima obtained from adaptive and gradient walks, respectively.

n	k_{pred}		$f_{\text{l.o.}}$		$f^{\star}_{\text{l.o.}}$		$f^{\star\star}_{\text{l.o.}}$		$f_{\text{pred}} = \eta$	
GC										
20	6.58	6.99	1.41	1.85	1.99	2.57	2.15	2.69	0.73	0.72
25	6.74	–	1.66	–	2.57	–	2.68	–	0.73	–
30	7.29	7.64	1.76	1.99	3.13	3.42	3.21	3.49	0.71	0.71
35	7.55	–	1.62	–	3.52	–	3.58	–	0.71	–
40	7.89	7.85	1.68	2.50	3.75	3.88	3.75	3.96	0.70	0.70
50	8.64	8.94	2.14	2.91	4.17	4.41	4.10	4.48	0.69	0.70
60	8.73	7.98	2.04	–	4.53	4.88	4.48	4.92	0.68	0.68
GCAU										
30	3.14	3.55	1.95	–	5.62	6.69	6.27	7.38	1.10	1.07
40	2.86	3.62	–	–	6.53	7.83	–	8.54	1.13	1.07
50	2.51	3.08	–	–	7.17	8.54	8.02	9.39	1.16	1.11
60	2.17	3.38	–	–	7.83	9.37	–	10.13	1.19	1.08
70	2.29	2.29	–	–	8.73	10.18	9.51	10.97	1.18	1.12
80	2.29	–	–	–	9.19	–	–	–	1.18	–

could accumulate a representative random sample of local optima (referred to as random optima in sect. IV.5). This allows a direct comparison with the prediction from the model as seen from table 7, $f_{\text{l.o.}}$ is still in disagreement with f_{pred}.

We find that some scaling properties are in common. In particular, the linear dependence of the walk length on n and the exponential decrease of the probability for finding a local optimum $p_{\mathrm{l.o.}}$ as a function of n. Estimates for the latter can already be derived by assuming that in highly correlated landscapes there should be $\mathcal{O}(1)$ local optima in a patch of radius λ [8] which after some calculations leads to

$$\log p_{\mathrm{l.o.}} = c_0 - n \cdot [\log 2 + \log(\kappa - 1)\alpha - 2(\alpha - 1/2)^2] \tag{37}$$

Table 8: Scaling of the number of local optima and walk length for for **GC** and **GCAU** landscapes. Shown are the slopes of linear fits to the calculated data (table 7). n-k model refers to the prediction from section V. Also shown are the rough estimates obtained from equations (37) and (38). Small letters refer to the older parameter set from Salser [22].

	$\log p_{\mathrm{l.o.}}/n$		L_{Adap}/n		L_{Grad}/n	
			GC			
GC landscape	-0.15	-0.12	0.15	0.18	0.10	0.12
$n-k$ model	-0.25	-0.25	0.25	0.25	0.18	0.18
equ.(37,38)	-0.35	-0.34	–	–	0.09	0.08
			GCAU			
GCAU landscape	< -0.3	≈ -0.2	0.44	0.32	0.26	0.20
$n - k$ model$hfil$	-0.60	-0.67	0.60	0.67	0.32	0.36
equ.(37,38)	-0.87	-0.94	–	–	0.26	0.30

By the same reasoning the length of a gradient walk should be roughly

$$L_{\mathrm{grad}} \approx \lambda \approx \alpha \cdot n. \tag{38}$$

Table 8 lists the scaled lengths of adaptive and gradient walks, as well as the logarithm of the probability of randomly hitting an optimum. The measured quantities are compared with the predictions from the n-k model and the crude estimates based on equations (37) and (38). The latter agree fairly well in the case of gradient walks with the observed data. The n-k model is at best in the ball-park. The ratios $L_{\mathrm{adap}}/L_{\mathrm{grad}}$, equation (32), are pretty close to the predictions. For the **GC** alphabet we find 1.45 and for **GCAU** we find 1.57, versus a prediction of 1.45 and 1.85, respectively.

VII Conclusions

Among the most important steps in understanding evolutionary adaptation is the construction of a model landscape based on the proper abstractions of the adapting entities. In this paper we explored in detail the statistics of a realistic and biologically motivated landscape induced by RNA folding. We compare its features with a widely used simple statistical model for rugged landscapes, known as the n-k model.

RNA energy as well as structure landscapes were explored by numerically computing landscape autocorrelation functions, random walk autocorrelation functions and density surfaces as a function of the nucleotide alphabet and the sequence length. The main results can be summarized as follows:

RNA Landscapes

1. The energy as well as the structure autocorrelation function is characterized to a reasonable approximation by one characteristic length scale.
2. The characteristic length of the energy and of the structure landscape scale linearly with sequence length.
3. The characteristic length for structures strongly depends on the nucleotide alphabet.

(i) Binary sequences, **AU** or **GC**, have very short correlation lengths, indicating that they are very likely to change their structure with few changes in the underlying sequences.

(ii) **GCXK** sequences, with **XK** denoting two artificial nucleotides with the same pairing strength as **GC**, are less sensitive to changes than binary sequences.

(iii) Natural **GCAU** sequences are even less sensitive than **GCXK**. We have checked the influence of the non-Watson-Crick pair **GU**. Disabling **GU** pairs in **GCAU** sequences strongly influenced the energy autocorrelation (shorter correlation length), but had no or little effect on the structure autocorrelation. We conclude that the sensitivity difference between **GCAU** and **GCXK** is due to the unequal base stacking and pairing energies associated with **GC** and **AU** pairs.

(iv) **ABCDEF** sequences (**GC** pairing strength) have a very low sensitivity.

This suggests that a natural four letter **GCAU** alphabet is a good compromise between (a) enough structural variety to support biological function, and (b) sufficient, but not excessive, stability towards changes in the sequence.

4. Exploration of the structure density surface suggests that there is a small region (compared to the diameter of the sequence space) around any random sequence, such that the sequences within that region fold into almost all minimum free energy structures.

Comparison with the n-k Model

The n-k model is a powerful and flexible, yet simple, tool for generating scalar landscapes with prescribed correlation structure. Although both RNA and n-k landscapes share some simple scaling behavior, they do not agree in important details concerning the statistics of local optima and the length of adaptive and gradient walks. We trace the disagreement between the fine structure of the two landscapes back to one basic difference. As detailed in section IV.5 the distribution of energy increments upon changes in one position are not properly described by a Gaussian in both binary alphabets and the natural alphabet. This is mainly due to a very high degree of neutrality, that is: neighboring sequences with identical minimum free energy and possibly, but not necessarily identical structure. With respect to energies there are only a few thousand different values [30]. In the n-k model all values are pairwise distinct with probability one. Even a discretization of the n-k model (e.g. cutting off all decimal places) would still remain Gaussian, without yielding a neutral neighbor peak of the kind observed in RNA folding. The physical process of polynucleotide folding — as far as it is properly abstracted by the presently used algorithm — is not in the class defined by the n-k model. The neutrality issue has profound effects on the number and the distribution of local optima as well as biased walks on both landscapes. These are the features in which the disagreement is most apparent. At the same time these are also the features that are the most relevant to evolutionary optimization.

The RNA energy landscape agrees to a good approximation with the n-k model in the basic functional form of the autocorrelation function (roughly a single decaying exponential). This is the basis for extracting a $k \approx 7$ to 8 as the number of context sites influencing the energetic contribution of each position, independently of sequence length. This coincides roughly with the typical size of secondary structure elements [5], and one may speculate about the nature of this coincidence.

Acknowledgements

This work was partly supported by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, project nos. S5305-PHY and P8526-MOB. The Santa Fe Institute is sponsored by core funding from the U.S. Department of Energy (ER-FG05-88ER25054) and the National Science Foundation (PHY-8714918). Pedro Tarazona acknowledges support from the *Direccion Central de Investigacion Cientifica y Tecnica*, project no. PB91-090.

Appendix

Derivation of the autocorrelation function for the n-k model

In case of the PR version the chance that a site energy remains unchanged when a single digit is flipped equals $1-\frac{k+1}{n}$, independent of whether other bits have been flipped before. Therefore,

$$1 - h_{PR}(d) = (1 - \frac{k+1}{n})^d. \qquad (A.1)$$

The RN version is only slightly more involved. A site energy f_i is unchanged if it is not one of the d positions that have been changed previously, and if it is not one of the k neighbors of any of the flipped sites. These events are statistically independent, and thus

$$1 - h_{RN}(d) = (1 - \frac{d}{n}) \cdot (1 - \frac{k}{n-1})^d. \qquad (A.2)$$

For the AN model, however, the situation is more complicated. First note that $h(1) = \frac{k+1}{n}$ since for a single bit flip it does not matter wether the assignment of neighbours is random or not, because the contributions to the site energies are pairwise independent. Now suppose that the $d \geq 2$ changes are labelled s_i such that

$$s_1 = 1 < s_2 < s_3 < \ldots < s_d. \qquad (A.3)$$

Evidently there are $\binom{n-1}{d-1}$ such flipping patterns. Let us now calculate the probability that a particular pair of flips occured ℓ sites apart, i.e. $s_{i+1} - s_i = \ell$. Suppose s_1 and s_2 are chosen this way. The remaining $d-2$ flips have therefore to be arranged within the remaining $n-\ell-1$ sites; there are $\binom{n-\ell-1}{d-2}$ such arrangements. Therefore we obtain the probability that two flips are separated by ℓ digits along the chain by

$$\phi_d(\ell) = \begin{cases} \binom{n-\ell-1}{d-2} / \binom{n-1}{d-1} & 1 < \ell < n-1 \quad \text{and} \quad 2 \leq d \leq n-\ell+1 \\ 0 & \text{otherwise} \end{cases}$$

If $\ell \leq k$ then ℓ site energies are changed as a result of changing the corresponding position. Otherwise all $k+1$ sites are changed. Thus the probability that a given site energy is changed by moving to Hamming distance d is

$$h_{AN}(d) = \frac{d}{n} \left\{ \sum_{\ell=1}^{k} \phi_d(\ell) \cdot \ell + (k+1) \left[1 - \sum_{\ell=1}^{k} \phi_d(\ell) \right] \right\}$$

References

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

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

[3] Weinberger E.D. Correlated and Uncorrelated Landscapes and How to Tell the Difference. *Biol. Cybern.* **63**,3 25-336 (1990).

[4] 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).

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

[6] Weinberger E.D. and P.F. Stadler. Why Some Fitness Landscapes are Fractal. Submitted to *J. Theor. Biol.* 1992.

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

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

[9] Stadler P.F. Corrlation of Landscapes of Combinatorial Optimization Problems Submitted to *Europhys. Lett.* (1992).

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

[11] Weinberger, E.D. $\mathcal{NP}$ Completeness of Kauffman's N-k Model, a Tuneably Rugged Fitness Landscape. Submitted to *Acta Appl. Math.* (1991).

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

[13] Mézard M. and G. Parisi. Mean-Field Equations for the Matching and Travelling Salesman Problems. *Europhys. Lett.* **2**, 913-918 (1986).

[14] Fu Y. and P.W. Anderson. Application of Statistical Mechanics to $\mathcal{NP}$-complete problems in combinatorial optimization. *J. Phys. A* **19**, 1605-1620 (1986).

[15] Bernasconi J. Low autocorrelation binary sequences: Statistical Mechanics and configuration space analysis. *J. Physique* **48**, 559-567 (1987)

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

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

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

[19] Waterman M.S. and T.F. Smith. RNA Secondary Structure. A Complete Mathematical Analysis. *Math. Biosc.* **42** (1978) 257-266.

[20] Zuker M. and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acid Res.* **9**,133-148.

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

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

[23] 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).

[24] Jaeger J.A., D.H. Turner and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA* **86**, 7706-7710 (1989).

[25] Hofacker I., P. Schuster and P.F. Stadler. Combinatorics of RNA Secondary Structures Submitted to *J. Comb. Theory.*

[26] Konings D.A.M. and P. Hogeweg. Pattern Analysis of RNA Structures. *J. Mol. Biol.* **207**, 597-614 (1989)

[27] Hogeweg P. and P. Hesper. Energy directed folding of RNA sequences. *Nucl. Acid Res.* **12**, 67-74 (1984).

[28] Shapiro B.A. An algorithm for comparing multiple RNA secondary structures. *CABIOS*, **4**, 381-393 (1988).

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

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

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

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

[33] Weinberger, E.D. A More Rigorous Derivation of Some Properties of Uncorrelated Fitness Landscapes. *J. Theor. Biol.* **134**, 125-129 (1988).

[34] Macken, C.A. and A.S. Perelson. Protein Evolution on Rugged Landscapes. *Proc. Natl. Acad. Sc.* **86**, 6191-6195 (1989).

[35] Macken C.A., P.S. Hagan and A.S. Perelson. Evolutionary Walks on Rugged Landscapes. *SIAM J. Appl. Math.* **51**, 799-827 (1991).

[36] Weinberger E.D. Local Properties of the Nk Model, a Tunably Rugged Landscape *Phys. Rev. A* **44**, 6399-6413 (1991).