Statistics of Landscapes Based on Free Energies, Replication and Degradation Rate Constants of RNA Secondary Structures †

By

Walter Fontana, Thomas Griesmacher, Wolfgang Schnabl,

Peter F. Stadler ‡ and Peter Schuster *

Institut für Theoretische Chemie der Universität Wien,

Los Alamos National Laboratory, and Santa Fe Institute N.M., USA

Submitted to Mh.Chem.

† Dedicated to Prof.Dr.Dr.h.c.mult.Viktor Gutmann

‡ Present Address: MPI für biophysikalische Chemie, Göttingen, Germany

* Mailing Address: Prof. Peter Schuster, Institut für Theoretische Chemie, Universität Wien, Währingerstraße 17, A 1090 Wien, Austria

Phone: (**43 1) 43 61 41 77

Fax: (**43 1) 40 28 525

E-Mail: pks@itc.univie.ac.at or

a8441dam@awiuni11.bitnet

Summary.

RNA secondary structures are computed from primary sequences by means of a folding algorithm which uses a minimum free energy criterion. Free energies as well as replication and degradation rate constants are derived from secondary structures. These properties can be understood as highly sophisticated functions of the individual sequences whose values are mediated by the secondary structures. Such functions induce complex value landscapes on the space of sequences. The landscapes are analysed by random walk techniques, in particular autocorrelation functions and correlation lengths are computed. Free energy landscapes were found to be of AR(1) type. The rate constant landscapes, however, turned out to be more complex. In addition, gradient and adaptive walks are performed in order to get more insight into the complex structure of the landscapes.

Keywords.

RNA secondary structures; RNA free energies; Value landscapes; Autocorrelation functions; Correlation lengths.

Zusammenfassung.

RNA Sekundärstrukturen werden aus den Primärsequenzen mit Hilfe eines Computeralgorithmus berechnet, welcher einem Kriterium minimaler freier Energien folgt. Freie Energien, Replikations- oder Abbaugeschwindigkeitskonstanten werden aus den Sekundärstrukturen berechnet. Man kann daher diese Eigenschaften als komplizierte Funktionen der Sequenzen aufgefassen, deren Zahlenwerte durch Vermittlung der Sekundärstrukturen erhalten werden. Diese Funktionen induzieren hochkomplexe Bewertungslandschaften im Raum der Sequenzen. Die Landschaften werden mit Hilfe von Irrflugtechniken analysiert. Im einzelnen werden Autokorrelationsfunktionen und Korrelationslängen berechnet. Die freien Energie-Landschaften sind vom AR(1) Typ. Die von den Reaktionsgeschwindigkeitskonstanten abgeleiteten Landschaften stellten sich hingegen als komplexer heraus. Zusätzlich werden die Bewertungslandschaften auch noch mit Hilfe von *Gradient* und *Adaptive Walks* untersucht, um mehr Einblick in ihre komplexe Struktur zu gewinnen.

1. Introduction

Statistical properties of ensembles of random biopolymers became an issue of current interest since new experimental techniques aim to exploit the enormous capabilities inherent in such ensembles [1−4]. Methods combining random assembly of biopolymers with selection represent a powerful counterpart to rational design. Another new trend in biotechnology produces variation by mutation and makes also use of selection to extract RNA molecules or proteins with the desired properties from organized mutant distributions. Selection is commonly introduced via replication under constraints provided by means of an appropriate setup or essay [5, 6]. The first *in vitro* selection experiments on RNA molecules were performed in the late sixties [7] and the remarkable capacities inherent in applied molecular evolution were discussed already some years ago [8, 9], but successful experimentation required tools which became available only recently. A third approach in biotechnology uses a natural device to carry out both processes of the evolutionary approach, randomization and selection: novel catalytic proteins are produced as antibodies by the immune system [10, 11].

Optimization based on natural or artificial selection is now commonly viewed as an adaptive walk on a highly complex *rugged landscape*. The concept of a fitness landscape is due to Sewall Wright [12]. For a more recent use of landscapes in evolution see [13−15]. Quantitative studies were almost exclusively performed by means of random model landscapes which were derived originally from spin glass theory. The *N*-*k* model conceived by Stuart Kauffman [14] represents the most intensively studied example. It can also serve as an appropriate reference system. Exceptions are investigations of value landscapes with a realistic biophysical background [16−18]. These studies are based on RNA folding into secondary structures. Mutant distributions in stationary replicating ensembles – denoted and characterized as *quasispecies* [19−21] – generally reflect the structure of the underlying value landscapes and one may use populations as experimental probes for the determination of the landscape structure [6, 22]. At the present state of the art the experimental approach yields important hints, but it is limited to few special cases and cannot provide global results on the properties of landscapes.

All quantitative values which are plotted in value landscapes are derived from spatial structures of biopolymers or even more complex entities involving ensembles of molecules. Any quantitative representation of realistic value landscapes has to deal therefore with the notoriously difficult problem to predict the structures of biopolymers from sequence data. Among the various attempts to compute structures of proteins or RNA molecules only the calculations of RNA secondary structures yield satisfactory results [23] at present. It was obvious therefore to choose this case for our studies.

What are the questions that can be answered by computation of statistical ensembles of RNA molecules? Distributions of thermodynamic and kinetic properties of structures, as for example free energies, replication and degradation rate constants, as well as their dependencies on chain lengths ν are of primary interest. How likely is it that closely related sequences have similar secondary structures and properties? What is the frequency of occurrence as well as the size and length distribution of the various structural elements, such as loops, stems, joints, and free ends? How do the answers to these questions depend on chain lengths, base compositions, and base alphabets? By base composition we mean the frequencies at which the four bases **G**, **A**, **C** and **U** occur in a particular RNA sequence. The base alphabet refers to the number and nature of the bases. In this paper we shall exclusively deal with thermodynamic and kinetic properties. The questions directly related to secondary structures and their elements will be dealt with in a forthcoming paper [24].

2. RNA secondary structures and value landscapes

The process of folding the *primary* sequence of an RNA molecule – to be understood as a finite string of symbols chosen from an alphabet of κ letters – into a three-dimensional *tertiary* structure can be partitioned into two steps:

1. folding of the string into a *quasi-planar* – two-dimensional – *secondary* structure by formation of complementary base pairs, **G**≡**C** or **A**=**U**, respectively, and
2. formation of a three-dimensional spatial structure from the quasi-planar folding pattern.

Secondary structure formation is modelled much more easily than their transformation into tertiary structures. At present there are no theoretical models available which can predict tertiary structures reliably. An additional problem is related to structure storage and handling: three-dimensional structures are very hard to encode in compact form – commonly Cartesian coordinates of all atoms have to be stored. As opposed to the difficulties in predicting and encoding tertiary structures, the computation of RNA secondary structures is much simpler and more reliable. Secondary structures – as we shall see – are readily encoded, stored and fairly easy to compare. This is mainly a consequence of the fact that the intermolecular forces stabilizing RNA secondary structures – base pairing and base pair stacking – are much stronger than those involved in three-dimensional structure formation. The dominant role of RNA secondary structures is also well documented in nature by the conservation of secondary structure elements in evolution[25–27].

Several assumptions are built into the folding algorithms through the definitions of RNA secondary structure:

- the secondary structure is a strictly planar graph – it contains no knots and hence crossing of strands can always be disentangled by rotation of partial structures,

Fig. 1: The sequence space of binary sequences of chain lengths $\nu = 3$ and $\nu = 5$. The sequence space is a point space in which every sequence is represented by one point. The points corresponding to sequences with Hamming distance $d = 1$ are connected by a straight line. The object obtained in that manner is a hypercube of dimension ν. RNA sequences in which two complementary bases out of the four natural ones are missing represent an example of binary sequences: **G**,**C** or **A**,**U**.

A

Fig. 2: Different representations of RNA secondary structures. The comparison shows two binary (**GC**) sequences **A** and **B**, their secondary structures in the compact encoded version g_k, the full graphical representation G_k ($5'$-ends are denoted by arrows, $\longrightarrow$) and the tree representation Γ_k. In the tree representation full circles ($\bullet$) denote base pairs, open circles ($\circ$) stand for unpaired bases and the root is symbolized by a full square ($\blacksquare$). Details are given in section 2. Note that the two sequences differ only in one position marked by $\Downarrow$.

- pseudoknots are considered as elements of tertiary structures, and
- RNA secondary structures can be partitioned into elements which contribute additively to thermodynamic and kinetic properties.

All non-additive contributions are assumed to be fairly small, and hence they can be attributed to the tertiary structure. Structural elements of secondary structures are:

1. **stems** or **stacks**, which represent the double helical regions of the structure,
2. **loops** and **bulges** consisting of internal unpaired bases,
3. **joints**, which are stretches of unpaired bases joining freely movable substructures, and
4. **free ends**.

Nucleotides in joints and free ends are also termed external. Several algorithms are available which allow to compute secondary structures from RNA sequences. The most widely used among them is based on dynamic programming and computes the minimum free energy secondary structure from the sequence [28]. Derivative algorithms allow to compute also suboptimal foldings [29, 30] and partition functions [31].

RNA sequences like other biopolymers are objects of combinatorial complexity, for example

$$\mathbf{I} \quad = \quad \{\textsf{\textbf{AUGCGUUGGACGAUGCAGUGAAACG}}\ldots\textsf{\textbf{GUAACG}}\}$$

Consequently, we are dealing with κ^ν different sequences of length ν. These numbers are enormous and taking statistically representative samples is not a simple task. The concept of a *sequence space* is very useful for representing the ordering of sequences. The sequence space is a discrete vector space. It was originally invented in information theory [32]. The Hamming distance

$$d_{ij} \quad = \quad d(\mathbf{I}_i, \mathbf{I}_j) \ ; \quad \mathbf{I}_i, \mathbf{I}_j \ \ldots \text{strings of chain length } \nu \tag{1}$$

counts the number of positions in which the two sequences $\mathbf{I}_i$ and $\mathbf{I}_j$ differ. It forms a metric on the sequence space. The sequence space of binary (**G**,**C**) sequences is simply a hypercube of dimension ν (figure 1). Sequence spaces of four-letter sequences are more sophisticated objects and we dispense here with a detailed discussion. There are several examples of applications of the concept of sequence space to problems in biophysics and biology [21, 33].

Folding the primary sequence of an RNA molecule into its most stable secondary structure may be represented formally by the expression

$$G_k \quad = \quad \mathcal{G}(\mathbf{I}_k) \ ; \quad k = 1, 2, \ldots, \kappa^\nu \quad , \tag{2}$$

where $\mathcal{G}$ is a function which stands for the folding algorithm. Based on the partitioning into structural elements RNA secondary structures may be encoded in compact form. As shown in figure 2 a secondary structure is encoded by assigning a lower case letter to every base of the primary sequence. The position of a given base in the sequence is the same in $\mathbf{I}_k$ and in the encoded secondary structure g_k. In particular we have:

1. stems, encoded by $aaa\cdots$, $bbb\cdots$, $ccc\cdots$, $ddd\cdots$, etc.,
2. loops, encoded by $xxx\cdots$,
3. joints, encoded by $yyy\cdots$, and
4. free ends, encoded by $zzz\cdots$.

The use of letters is not arbitrary here: the later the letters come in the alphabet, the more flexible are the corresponding parts of the RNA molecule. Stem regions are more rigid than loops, loops in turn are more rigid than joints and joints are less flexible than the free ends. It is not necessary to distinguish between different loops or joints in the encoded notation g_k since they can be reconstructed from the stacks whose positions are specified by denoting the stacks from the $5'$- to the $3'$-end in their sequence of occurence.

A very convenient and versatile way to represent and compare RNA secondary structures uses the concept of trees. A coarse tree representation of secondary structures was already suggested by Michael Zuker and David Sankoff in their well known paper on the folding algorithm [28]. It was resumed and used for comparisons of secondary structures on a large scale by Bruce Shapiro [34, 35]. As shown in figure 2 the tree representation,

$$\Gamma_k \quad = \quad \mathcal{T}(G_k) \ , \tag{3}$$

can be easily extended such that it covers all structural details: every base pair is represented by an internal node, every single unpaired base by an external node or *leaf*. We add one more node which does not correspond to a physical unit or the RNA molecule as the *root* of the tree in order to avoid that simple secondary structures with free ends are represented by *forests*. This representation makes explicit that the folding process can be viewed as a map between linear and nonlinear combinatorial structures: sequences and trees. Tree representations of secondary structures provide one important advantage: a distance between secondary structures with the properties of a metric,

$$\Delta_{ij} \quad = \quad \Delta(\Gamma_i, \Gamma_j) \quad , \tag{4}$$

is obtained by a tree editing procedure [36, 37]. Paulien Hogeweg and Danielle Konings conceived an alternative graphical method for the comparison of RNA secondary structures called *mountain representation* [38−40].

A *value landscape* is obtained by taking the hypercubical sequence space as the support of a function that assigns a value to every sequence. The conventional value to be plotted in biology is the fitness of the phenotype that is replicated. The analogy to concepts in biology may be carried further: primary RNA sequences of $\mathbf{I}_k$ are equivalent to genotypes and the thermodynamically most stable secondary structures, $G_k = \mathcal{G}(\mathbf{I}_k)$ are the analogues of phenotypes. The process of folding the one-dimensional string of symbols into the two-dimensional secondary structure is tantamount to the "unfolding of the phenotype". The biological concept of a fitness landscape is generalized to a value landscape in which an arbitrary scalar property of the phenotyp is plotted over sequence space. In particular we shall deal here with landscapes derived from thermodynamic and kinetic properties of RNA molecules in their most stable secondary structures.

3. Autocorrelation functions of time series

The correlation coefficient of two random variables X_i and X_k is defined by

$$\begin{aligned}\rho_{ik} &= \frac{\mathrm{cov}\,(X_i, X_k)}{\sqrt{\mathrm{var}\,(X_i)\,\mathrm{var}\,(X_k)}} \\ &= \frac{\big\langle \big(X_i - < X_i >\big)\big(X_k - < X_k >\big)\big\rangle}{\sqrt{\big\langle \big(X_i - < X_i >\big)^2\big\rangle\big\langle \big(X_k - < X_k >\big)^2\big\rangle}} \ .\end{aligned} \tag{5}$$

Throughout this paper we use the notation $\langle \cdot \rangle$ for expectation values, cov is used for the covariance matrix and $\mathrm{var}\,(X_i) \equiv \mathrm{cov}\,(X_i, X_i)$ denotes its diagonal terms, the variance of the random variable X_i. The correlation coefficient apparently fulfils: $-1 \leq \rho_{ik} \leq 1$. Let us assume that both random variables, X_i and $X_k = X_{i+k}$ are created by the same stochastic process yielding the time series

$$\mathcal{X}_t = \big(X_0, X_1, X_2, \ldots, X_i, X_{i+1}, \ldots, X_{i+k}, \ldots\big) \ . \tag{6}$$

X_{i+k} is the value of the random variable $\mathcal{X}_t$ exactly k time steps later than X_i. We assume stationarity of the stochastic time series. This condition is readily casted in the form:

$$\big\langle X_i^r\, X_j^s \big\rangle = \big\langle X_{i+k}^r\, X_{j+k}^s \big\rangle \quad \text{for all} \quad i, j, k, r, s \ .$$

The correlation coefficient may now be expressed as a function of k thus leading to the autocorrelation function

$$\rho(k) = \frac{\big\langle \big(X_i - < X_i >\big)\big(X_{i+k} - < X_i >\big)\big\rangle}{\big\langle \big(X_i - < X_i >\big)^2\big\rangle} \ . \tag{7}$$

The denominator becomes simpler because of the stationarity condition: expectation values, variances and all higher moments of variables with time offset become identical: $\langle X_i \rangle = \langle X_{i+k} \rangle$, $\mathrm{var}\,(X_i) = \mathrm{var}\,(X_{i+k})$, etc.

Two variants of equation (7) are important for the issues pursued here and in forthcoming papers on RNA value landscapes:

$$\rho(k) = 1 - \frac{\langle X_i^2 \rangle - \langle X_i X_{i+k} \rangle}{\langle X_i^2 \rangle - \langle X_i \rangle^2} \ . \tag{7a}$$

This equation is useful for numerical computation and makes the limits of the autocorrelation function easily intelligible: $\lim_{k\to 0}\rho(k) = 1$, and $\lim_{k\to\infty}\langle X_i X_{i+k}\rangle = = \langle X_i\rangle\langle X_{i+k}\rangle = \langle X_i\rangle^2$, since the two random variables X_i and X_{i+k} become independent for sufficiently large k. Eventually we find $\lim_{k\to\infty}\rho(k) = 0$.

Let us assume that X_j is independent of X_i and X_{i+k} but nevertheless chosen randomly from the same time series $\mathcal{X}_t$. Then the autocorrelation function can also be written as [41]

$$\rho(k) \;=\; 1 - \frac{\langle (X_i - X_{i+k})^2\rangle}{\langle (X_i - X_j)^2\rangle} \; . \tag{7b}$$

This equation allows to define an autocorrelation function also in cases where only distances between objects are meaningful. An example is given by the distances between RNA secondary structures (see section 2),

$$X_i - X_j \;\Longrightarrow\; X_{ij} \;=\; \Delta(\Gamma_i, \Gamma_j)$$

that will be analysed in detail in a forthcoming paper [41].

In the simplest cases autocorrelation functions have the shape of decaying exponentials, $\rho(k) = \exp(-\lambda k)$. They correspond to time series which are created by AR(1) processes [42]. The reciprocal value of the decay constant is called the correlation length $\ell = \lambda^{-1}$. It represents the value of k at which the autocorrelation function has dropped to e^{-1}. Value landscapes on which random walks lead to AR(1) processes may be characterized as AR(1) landscapes. In particular random landscapes as obtained from Kauffman's N-k model have this property [43]. Other examples of AR(1) landscapes are the Sherrington-Kirkpatrick spin-glass [44] and the symmetric traveling salesman problem [45].

4. Statistical evaluation of value landscapes

Value landscapes may be explored by random walks in sequence space [43]. At first we consider the free energy autocorrelation function. A series of RNA molecules is created by successive random single base exchanges:

$$\mathbf{I}_0 \to \mathbf{I}_1 \to \mathbf{I}_2 \to \ldots \to \mathbf{I}_i \to \mathbf{I}_{i+1} \to \ldots \to \mathbf{I}_{i+k} \to \ldots \quad . \tag{8}$$

The random walk then occurs along the edges of the hypercube (figure 2) – or along those of a more complex object in case of four-letter sequences, since subsequent molecules have always Hamming distance one: $d(\mathbf{I}_1, \mathbf{I}_0) = d(\mathbf{I}_2, \mathbf{I}_1) = \ldots = d(\mathbf{I}_{i+1}, \mathbf{I}_i) = 1$. Planar secondary structures $G_i = \mathcal{G}(\mathbf{I}_i)$ are computed for all RNA molecules by means of the algorithm described in section 2. The free energies of the RNA molecules in their most stable (0°K) structures, $f_i = \mathcal{F}(G_i)$, are readily obtained from the structure computation. In essence, the series (8) is mapped onto a – time ordered – set of numbers:

$$f_0 \to f_1 \to f_2 \to \ldots \to f_i \to f_{i+1} \to \ldots \to f_{i+k} \to \ldots \tag{9}$$

The random walk in sequence space thus yields a stochastic process on the free energy landscape that can be analysed with the statistical techniques outlined in the previous section 3.

RNA secondary structures – once obtained by the folding algorithm – can be used to compute other quantities of interest. As opposed to free energies, the kinetic constants of replication, a_k, and degradation, d_k, cannot be computed straightway from known secondary structures G_k. There are no satisfactory models available which are based exclusively on the knowledge in biophysical chemistry. We use therefore a very crude estimate of these quantities. It is well known from virus specific RNA replication by Qβ replicase that only single stranded molecules are accepted as templates. The secondary structure has to *melt* in order to make replication possible. An estimate of the rate constant of melting may be used therefore as a simple model for the replication rate constant. The expression given below takes care of the cooperativity in the melting process. Degradation on the other hand can be modelled by taking into account all possible attacks of a hydrolytic agent or an enzyme with nuclease activity on the single stranded regions of the secondary structure G_k. We present here two examples of such model equations, one for replication rate constants [16, 17],

$$a_k = \mathcal{A}(G_k) = \alpha_0 - \alpha_1 \sum_{j=1}^{s^{(k)}} \frac{n_j^{(k)}(1+n_j^{(k)})^3}{(1+n_j^{(k)})^4 + \alpha_2} \; ; \quad k = 1,2,\ldots,2^\nu$$

and one for degradation rate

$$d_k = \mathcal{D}(G_k) = \beta_0 + \beta_1 \sum_{j=1}^{u^{(k)}} \frac{u_j^{(k)}}{u_{max}} \exp\Big\{(u_j^{(k)} - u_{max})/u_{max}\Big\} + \frac{\beta_2}{\nu} z^{(k)} \; .$$

The two equations contain six empirical parameters: α_0, α_1, α_2, β_0, β_1 and β_2. The other quantities are related to the secondary structure G_k. By $n_j^{(k)}$ we denote the number of base pairs in the j-th stack of the secondary structure G_k, $s^{(k)}$ is the number of stacks in this structure. In the second equation the number of bases in the j-th loop of the secondary structure G_k is denoted by $u_j^{(k)}$. This structure has $u^{(k)}$ loops and there is a maximum loop size u_{max} above which loops are considered as completely mobile elements like free ends. The total number of bases in large loops, joints and free ends is given by $z^{(k)}$. Both expressions were used as model equations in a computer simulation of optimization of RNA secondary structures by mutation and selection [16, 17].

Depending on the quantity plotted on the value landscape we distinguish autocorrelation functions of free energies, $\rho_f(k)$, of replication rate constants, $\rho_a(k)$, and of degradation rate constants, $\rho_d(k)$. Since the forthcoming considerations apply to all autocorrelation functions we drop the index. Let us first consider the nature of the random walk in sequence space. Although every step of the random

Fig. 3: The relation between the length of a random walk (s) and the Hamming distance between the first and the last sequence (d). For the purpose of illustration we choose binary sequences, $\kappa = 2$, and a walk length of $s = 5$. In this case we have the simple restriction that the sum of s and d has to be an even number. Thus, only the Hamming distances $d = 1, 3$ and 5 may occur.

walk is of Hamming distance one the walk length (s) need not coincide with the Hamming distance (d) between the first and the last sequence (figure 3). Indeed we find a probability distribution of Hamming distances covered by a walk of length s: $\phi_{ds}(\nu, \kappa)$ is the probability that a random walk of s steps ends at a sequence with Hamming distance d from the starting sequence. It depends on the chain length of the sequence, ν, and on the number of letters in the genetic alphabet, κ. The Hamming distance d can neither be larger than the walk length s nor can it be larger than the chain length ν and hence

$$\phi_{ds}(\nu, \kappa) \;=\; 0 \quad \text{if} \quad d > \min(s, \nu) \; . \tag{10}$$

The probabilities fulfil the conservation relation

$$\sum_{d=0}^{\min(\nu,s)} \phi_{ds}(\nu, \kappa) \;=\; 1 \; ,$$

since every walk of length s has to yield some Hamming distance d. The probability distribution $\phi_{ds}(\nu, \kappa)$ is eventually derived from the recursion

$$\begin{aligned} \phi_{ds}(\nu, \kappa) \quad = \quad & \phi_{d-1,s-1}(\nu, \kappa) \cdot \frac{\nu - d + 1}{\nu} \; + \\ & + \phi_{d,s-1}(\nu, \kappa) \cdot \frac{d}{\nu} \cdot \frac{\kappa - 2}{\kappa - 1} \; + \\ & + \phi_{d+1,s-1}(\nu, \kappa) \cdot \frac{d+1}{\nu} \cdot \frac{1}{\kappa - 1} \end{aligned} \tag{11}$$

with $\phi_{ds}(\nu, \kappa) = 0$ if $d < 0$, and the initial condition $\phi_{00}(\nu, \kappa) = 1$.

Some useful general relations are readily derived:

$$\phi_{ss}(\nu,\kappa) \quad = \quad \begin{cases} 1, & \text{if } s=0,1 \\ \frac{(\nu-1)(\nu-2)\ldots(\nu-s+1)}{\nu^{s-1}} = \frac{(s-1)!}{\nu^{s-1}}\binom{\nu-1}{s-1}, & \text{if } s>1 \ . \end{cases} \tag{12}$$

In the case of binary sequences ($\kappa = 2$) the recursion (11) simplifies to

$$\phi_{ds}(\nu,2) \ = \ \phi_{d-1,s-1}(\nu,2)\cdot\frac{\nu-d+1}{\nu} + \phi_{d+1,s-1}(\nu,2)\cdot\frac{d+1}{\nu}\ , \tag{11a}$$

and, in addition, every second element of the transformation matrix vanishes:

$$\phi_{ds}(\nu,2) \ = \ 0 \quad \text{if} \quad d+s=2j+1\ ;\ j=0,1,2,3,\ldots\ . \tag{11b}$$

The probability distribution is applied to compute the autocorrelation as a function of the Hamming distance, $p(d)$, from the autocorrelation as a function of the walk length, $\rho(s)$, with the latter obtained directly from the random walk:

$$\rho(s) \quad = \quad \sum_{d=0}^{\min(\nu,s)} \phi_{ds}(\nu,\kappa)\cdot p(d)\ . \tag{13}$$

For $s \leq \nu$ the desired autocorrelation function $p(d)$ is now readily computed by recursion from

$$\begin{aligned} p(0) \ &= \ \rho(0) \ = \ 1\ ,\ p(1) \ = \ \rho(1)\ , \quad \text{and} \\ p(s) \ &= \ \frac{1}{\phi_{ss}(\nu,\kappa)}\left(\rho(s) - \sum_{d=0}^{s-1}\phi_{ds}(\nu,\kappa)\cdot p(d)\right)\ ; \quad s=2,3,\ldots\ . \end{aligned} \tag{14}$$

In the case $s > \nu$ of the autocorrelation function p(d) vanishes since the Hamming distance can never be longer than the diameter of the sequence space (see also equation 10).

5. Numerical results from random walks

Sampling of RNA structures by means of random walks with step size one in Hamming distance seems to be a rather simple task. There are, however, fundamental and technical problems that have to be overcome:

- Sequence spaces are objects of combinatorial complexity, and taking statistically representative samples is a non-trivial problem already at moderate chain lengths ($\nu = 40$ and larger).

Consider for example natural (**GACU**) sequences of chain length $\nu = 40$. The sequence space comprises $4^{40} \approx 1.21 \times 10^{24}$ different sequences. Computational requirements become prohibitive for random walks with walk lengths of $s = 10^6$ steps or more. We found a workable solution to the problem which consists in the accumulation of short random walks each about 1000 steps long. Every walk is started anew by resetting the random number generator. Then a few 10^5 points were found to be sufficient to achieve convergence of the most important statistical quantities.

- Storage of the folded secondary structures provides a formidable memory problem.

Files containing the collection of secondary structures in the compact encoded form g_k commonly exceed several tens of megabytes and sufficient storage capacity is a prerequisite for this type of computations.

The calculations reported here were performed with the same parameter set as used in our previous studies [16, 17]. This choice was made for the sake of consistency. We mention, however, that new parameters are also available [46] which were derived from a largely extended empirical data base and hence allow more precise predictions.

Free energy, replication and degradation rate constant landscapes of RNA molecules derived from a two-letter alphabet (**GC**) as well as those of natural four

Fig. 4: Expectation values of free energies ($<f>$) of RNA molecules in their most stable secondary structures as functions of the chain length ν. The free energies of binary sequences (**GC**) are denoted by ▪. For those of natural four letter sequences (**GACU**) the symbol • is used. Standard deviations ($\sqrt{\mathrm{var}\,(f)}$) – multiplied by a factor 10 – are shown as ⋄ or as ∘, respectively.

Fig. 5: Expectation values of replication ($< a >$) and degradation rate constants ($< d >$) as well as excess productivities ($< e >$) of two letter (**GC**) RNA molecules in their most stable secondary structures as functions of the chain length ν. Replication rate constants of binary sequences are denoted by ▪. For degradation rate constants and excess productivities the symbols • or ∘ are used, respectively.

Fig. 6: Expectation values of replication ($< a >$) and degradation rate constants ($<d>$) as well as excess productivities ($<e>$) of four letter RNA molecules in their most stable secondary structures as functions of the chain length ν. Replication rate constants of natural four letter sequences (**GACU**) are denoted by ▪. For degradation rate constants and excess productivities the symbols • or ∘ are used, respectively.

letter sequences (**GACU**) were explored. The chain lengths were typically varied from $\nu = 20$ to $\nu = 140$. In figures 4, 5 and 6 we show the expectations values, $< f(\nu) >$, $< a(\nu) >$, $< d(\nu) >$ and $< e(\nu) >$ as functions of the chain length together with their standard deviations. Excess productivities, $e_k = a_k - d_k$ were included in the analysis because they represent the quantities which determine the outcome of selection in molecular evolution experiments [21].

At first we discuss and analyse the results for free energies. As shown in figure 4 the absolute values of mean free energies increase linearly with chain length. As expected **GC**-sequences are more stable than **GACU**-sequences since

Fig. 7: An analytical approximation to the autocorrelation function $\rho(s)$ for binary sequences ($\kappa = 2$) of chain length $\nu = 30$ as described in equation (15). The points at which the function is defined are denoted by $\bullet$. The straight line represents the autocorrelation function expressed in Hamming distance $p(d)$.

(1) **G≡C** base pairs are stronger than **A=U** base pairs, and

(2) **GC**-sequences form on the average more base pairs than **GACU**-sequences.

In previous papers we found that free energies in random samples of RNA secondary structures follow roughly a normal distribution [16, 17]. This implies that they are well characterized by standard deviations. The distribution of the energies of four letter sequences is wider – relative to the mean value – than those of **GC**-sequences and increases more strongly with the chain length ν. This implies that the secondary structures of the natural **GACU**-sequences are richer and more variable in the details than those of their two-letter counterparts.

Autocorrelation functions $\rho(s)$ were computed according to equation (7). They yield curves in $\log \rho(s)/s$-plots and hence are more complex than single exponential decay functions. A correction for the difference between walk length (s) and Hamming distance (d) according to the recursion (11) yields an autocorrelation function which can be fitted well by a single exponential (figure 7).

In the case of sequences from a two-letter alphabet ($\kappa = 2$) an analytic approximation can be derived for the transformation $\rho(s) \Longleftrightarrow p(d)$. This approximation assumes that the random walk in Hamming distance gives rise to an AR(1) process, $p(d) = \exp(-\lambda d)$, and neglects terms of order ν^{-2} and smaller. It is valid for $s, d \ll \nu$:

$$\rho(s) \;=\; \left(1 + \frac{s(s-1)}{2\nu}\left(e^{2\lambda} - 1\right) + \mathcal{O}\left(\nu^{-2}\right)\right) \cdot \exp(-\lambda s) \tag{15}$$

It is easily verified that the correction term accounts for the curvature observed in the $\log \rho(s)/s$-plot mentioned above (see figure 7).

Fig. 8: The autocorrelation function of the free energy $\rho_f(s)$ for binary (**GC**) sequences of chain length $\nu = 30$ as obtained by sampling 600 000 data points ($\bullet$). An additional sampling of the same series of sequences with respect to Hamming distances d [47] yields the autocorrelation function expressed in Hamming distance $p_f(d)$ in direct computation. The individual points ($\circ$) fulfil the linear relationship, $\log p(d) = -d/\ell$ with $\ell = 3.256$, to a very good approximation.

Table 1: Autocorrelation functions of free energies and their analytical approximations. Binary (**GC**)-sequences of chain length $\nu = 30$ are considered. The computed data were taken from a random walk of 600 000 steps. They were collected in 600 packages of 1000 points each and sampled with respect to walk length s or Hamming distance d, respectively.

	Walk length		Hamming distance		
d, s	$\rho(s)$ Sample	$\rho(s)$ Equ.(15)	$p(d)$ Sample	$p(d)$ Equ.(14)	$p(d)$ $\exp(-d/\ell)$
0	1	1	1	1	1
1	0.700	0.735	0.696	0.700	0.736
2	0.532	0.557	0.520	0.516	0.541
3	0.418	0.432	0.393	0.388	0.398
4	0.336	0.343	0.296	0.293	0.293
5	0.275	0.277	0.228	0.221	0.215
6	0.229	0.226	0.173	0.167	0.158
7	0.192	0.186	0.124	0.124	0.116
8	0.161	0.154	0.087	0.086	0.086
9	0.137	0.127	0.058	0.053	0.063
10	0.115	0.106	0.021	0.030	0.046

In order to verify equation (15) by comparison with a computed ensemble of

Fig. 9: The correlation length ℓ of free energies f for four letter (**GCAU**) sequences (■) and for two letter (**GC**) sequences (•) in their most stable secondary structures as a function of the chain length ν.

secondary structures we computed 600 000 (**GC**) sequences in 600 packages to 1000 sequences each and sampled with respect to walk length s and Hamming distances d [47]. This sample of free energies provides enough material for an appropriate test. Numerical data for the autocorrelation functions and their approximations are collected in table 1. As shown in figure 8 the autocorrelation function $p_f(d)$ can be approximated by a single exponential with very good agreement. There is no reason, however, that a random walk on an RNA free energy landscape should exactly meet the conditions of an AR(1) process and indeed careful inspection of the computer data shows small systematic deviations from the straight line. The nature of these higher order corrections will be analysed and discussed in a forthcoming paper on random walk studies with updated parameter sets [47].

Correlation lengths of free energies are considered now as functions of the chain length ν. In figure 9 we show plots of $\ell_f^{\mathbf{GC}}(\nu)$ and $\ell_f^{\mathbf{GACU}}(\nu)$. Two features are immediately evident:

- the correlation lengths of free energies increase roughly linearly with the chain length ν, and
- free energies of **GC** sequences have much shorter correlation lengths than those of sequences from the natural four-letter alphabet.

A closer inspection of the curves $\ell_f(\nu)$ shows a slightly sigmoid shape. This behaviour suggests the existence of a limit value of the correlation length for long sequences.

The expectation values of the model rate constants, $<a>$, $<d>$ and $<e>$, show non-monotonous dependence on the chain length ν (figures 5 and 6). For short sequences $<a(\nu)>$ increases with chain length and – after passing through

Fig. 10: The correlation length ℓ of free energies f ($\blacksquare$), replication rate constants a ($\bullet$), degradation rate constants d ($\diamond$) and excess productions e ($\circ$) of binary (**GC**) sequences in their most stable secondary structures as a function of the chain length ν.

Fig. 11: The correlation length ℓ of free energies f ($\blacksquare$), replication rate constants a ($\bullet$), degradation rate constants d ($\diamond$) and excess productions e ($\circ$) of four letter (**GCAU**) sequences in their most stable secondary structures as a function of the chain length ν.

a maximum – it decreases with increasing values of ν. The mean degradation rate constant, on the other hand, shows inverse behaviour: it starts from high values at short sequences, passes through a flat minimum and increases then with further increasing chain length ν. The two opposing dependencies on the chain length are readily interpreted by inspection of the corresponding equations in section 4: the replication rate constant is mainly determined by the number and the sizes of stacks, and the degradation rate constant reflects the number and the sizes of loops. It is easily verified that the number of loops equals the number of stacks and the size distributions become roughly alike by averaging over large samples. The dependence of $< e >=< (a-d) >$ on the chain length ν is derived readily and apparently parallels that of $< a >$. Mean rate constants computed for two-letter (**GC**) sequences show the same qualitative behaviour than those computed for four-letter (**GACU**) sequences. With the two-letter sequences, however, the extrema occur at substantially smaller chain length.

The models of rate constant landscapes were found to be more complex than those of the free energies: after transformation into $p(d)$ the autocorrelation functions are still curved and this indicates that the stochastic processes corresponding to the random walks are not AR(1). Here we shall not investigate the nature of these landscapes further. Since the curvature is only moderate we computed ap-

Fig. 12: Probability densities $P(s)$ of the length of gradient (**A**) and adaptive (**B**) walks on free energy landscapes. The data shown in the plot were computed for 20 000 walks involving **GC**-sequences of chain lengths $\nu = 30$ (•,○) and $\nu = 50$ (■,◇).

Fig. 13: Probability densities $P(s)$ of the length of gradient (•,■) and adaptive walks (○,◇) on free energy landscapes. The data shown in the plot were computed for 20 000 walks involving **GC**-sequences of chain lengths $\nu = 30$ (**A**) and $\nu = 50$ (**B**).

proximate correlation lengths by linear interpolation between the two points in the neighbourhood of $\log p(\ell) = -1$.

In figures 10 and 11 the approximate correlation lengths of the three rate constants a, d and e are plotted and compared with the correlation lengths of the free energies for **GC**- and **GACU**-sequences. In general the correlation lengths of the degradation rate constants d are much smaller than those of the free energies, and replication rate constant landscapes a, in turn, have smaller correlation length than the corresponding d-landscapes. As with the free energies the correlation lengths of **GC**-sequences are much shorter than those of their four-letter counterparts.

6. Adaptive and gradient walks on free energy landscapes

Two other classes of processes on value landscapes, gradient and adaptive walks, were studied in addition to random walks. These processes are created by series of sequences as shown in equation (8) whose successors have Hamming distance $d(\mathbf{I}_{i+1}, \mathbf{I}_i) = 1$. In contrast to random walks they are restricted to decreasing values of free energies in the series

$$f_0 > f_1 > f_2 > \ldots > f_i > f_{i+1} > f_{i+2} \ldots \quad . \tag{9a}$$

Fig. 14: Probability densities of stepwise improvements of free energies as obtained from 20 000 gradient (**A**) and 20 000 adaptive walks (**B**) on a free energy landscape of binary (**GC**) sequences of chain length $\nu = 30$. We show densities for the first and second steps as well as for the sum of all remaining steps ($s \geq 3$).

and end always in local minima of the free energy landscape. Consequently they provide direct information on the distribution of minima. We distinguish

(1) gradient walks – deterministic walks of step size Hamming distance one in which always the highest value of the free energy f in the nearest neighbourhood of the current sequence is chosen – and

(2) adaptive walks which choose at random a sequence in the nearest neighbourhood which fulfils equation 9a.

The results of four selected walks on two free energy landscapes derived from two-letter (**GC**) sequences of chain length $\nu = 30$ and $\nu = 50$ are shown in figures 12 and 13. In order to obtain statistically significant samples 20 000 walks were performed for each class.

In table 2 we summarize the numerical values of mean walks lengths and mean free energy gains. Starting from the same point in sequence space the adaptive walk is necessarily always longer or – in the limit – as long as the gradient walk. The results shown in figure 12 clearly demonstrate this fact. An increase in the chain length ν of the sequences manifests itself in longer walks (figure 13). Free energy landscapes of longer sequences thus have a smaller fraction of local minima. The fraction of vertices representing local minima of the free energy is obtained also directly as the fraction of adaptive or gradient walks having zero walk length (table 2). It amounts about 2% for **GC** sequences of chain length $\nu = 30$ and decreases by one order of magnitude if the chain length is raised to $\nu = 50$.

Intuitively we would expect adaptive walks to lead to deeper local minima than the gradient walks do. According to table 2 this might be fulfilled at longer

Table 2: Fraction of vertices corresponding to local minima of the free energy landscape, γ_m, mean walk lengths, $< s >$, and expectation values of the improvements in free energies, $<\Delta f>$ [kJ.mol^{-1}], for gradient and adaptive walks.

Chain length	Fraction of minima	Gradient walks		Adaptive walks	
ν	γ_m	$<s>$	$<\Delta f>$	$<s>$	$<\Delta f>$
30	0.0227	3.1	-16.1	4.4	-15.7
50	0.0027	5.5	-25.6	8.1	-25.9

chain lengths. For $\nu = 30$ we even observe an opposite trend indicating that the density of local minima is fairly high and thus the longer reach of the adaptive walk does not provide any advantage over the gradient walk.

Gains in free energies during adaptive or gradient walks can be resolved into probability densities for individual steps – examples are shown in figure 14. As expected the gradient walks show larger improvements than the adaptive walks in the first step whereas the opposite is true for later steps. The probability densities are rather complex and show series of high peaks. The corresponding preferred values of Δf are independent of the nature of the walk and reflect regularities of RNA secondary structures.

7. Conclusions

Free energy landscapes of four-letter (**GACU**)-sequences are substantially less rugged than those derived from two-letter (**GC**)-sequences. This is documented well by remarkably longer correlation lengths. The base composition of RNA sequences thus represents an interesting tool that allows to tune the structure of landscapes. Such a tool might well be important in evolutionary optimization.

The transition from **GC**- to **GACU**-sequences introduces two different, major changes into the logic and the physics of polynucleotide folding:

- two complementary base pairs instead of one make it more difficult to find matching counterparts to segments of the sequence and hence, global refolding of structures after the exchange of one, two or a few bases is less likely, and
- weaker interactions between complementary bases destabilize short stacks of base pairs, some of them do not fold at all and secondary structures become less complex.

Both effects change free energy landscapes in the same direction and it is not yet possible to separate the results of general features like the increase in the number of base pairs from specific effects of the **GACU**-system. This question will be addressed in forthcoming studies on free energy landscapes of RNA molecules and model polynucleotides.

Correlation length as well as the results of gradient and adaptive walks demonstrate that the free energy landscapes of longer sequences are smoother than those of their shorter homologues. The relative number of local minima illustrates the effect of increasing chain lengths best. On the average 230 out of 10 000 vertices are local minima of the free energy landscape of **GC** sequences with chain length $\nu = 30$. In increase in the chain length from $\nu = 30$ to $\nu = 50$ reduces the relative number of minima by about one order of magnitude to 27 minima out of 10 000 vertices.

The free energy landscapes can be approximated very well as AR(1) landscapes. This implies that a number of relations are fulfilled [43]. For example, the landscapes are statistically isotropic – this means that all vertices are equivalent for statistical analysis – and the free energies from randomly chosen samples of sequences are normally distributed.

The model landscapes for replication and degradation rate constants are more rugged and more complex. The correlation lengths are smaller and random walks lead to stochastic processes more complex than AR(1). It is premature to conclude that kinetic value landscapes will generally show higher complexity than their thermodynamic counterparts. Further studies on more realistic kinetic value landscapes, for example on one that describes *melting* of RNA secondary structures, are under way.

Acknowledgements

Financial support for the work reported here was provided by the *Jubiläumsfonds der Österreichischen Nationalbank* (project no.3819), by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (project no.P 6864), by the German *Volkswagen-Stiftung*, by the *John D. and Catherine T. Mac Arthur Foundation*, by the *National Science Foundation* (PHY-8714918) and by the U.S. *Department of Energy* (ER-FG05-88ER25054). Computer time on the IBM 3090 mainframe was generously supplied by the *EDV-Zentrum der Universität Wien*. Useful hints in stimulating discussions given by Professors Doyne Farmer, Stuart Kauffman, David Lane, John McCaskill, Alan Perelson, and Ed Weinberger are gratefully acknowledged. We thank Dr.Michael Ramek for providing TEX Macros for drawing diagrams.

References

[1] Horwitz, M.S.Z., Dube, D.K., Loeb, L.A. (1989) *Genome* **31** : 112

[2] Joyce, G.F. (1989) *Gene* **82** : 83

[3] Tuerk, C., Gold, L. (1990) *Science* **249** : 505

[4] Ellington, A.D., Szostak, J.W. (1990) *Nature* **346**: 818

[5] Husimi, Y., Keweloh, C. (1987) *Rev.Sci.Instrum.* **58**: 1109

[6] Biebricher, C.K. (1988) *Cold Spring Harbor Symp.Quant.Biol.* **52**: 299

[7] Spiegelman, S. (1981) *Quart.Rev.Biophys.* **4**: 36

[8] Eigen, M. (1986) *Chemica Scripta* **26B**: 13

[9] Kauffman, S.A. (1986) *J.theor.Biol.* **119**: 1

[10] Lerner, R.A., Tramontano, A. (1988) *Sci.Am.* **258/3**: 42

[11] Schulz, P., Lerner, R.A. (1991) *At the cross-roads of chemistry and immunology: catalytic antibodies.* Science, in press.

[12] Wright, S. (1932) *Procceedings of the Sixth International Congress on Genetics.* **1**: 356

[13] Kauffman, S.A., Levin, S. (1987) *J.theor.Biol.* **128**: 11

[14] Kauffman, S.A. (1989) *Adaptation on rugged fitness landscapes.* In: Stein, D., ed. Complex Systems. SFI Studies in the Science of Complexity. Addison-Wesely Longman, Redwood City (Cal.), pp.527-618.

[15] Macken, C.A., Perelson, A.S. (1989) *Proc.Natl.Acad.Sci.USA* **86**: 6191

[16] Fontana, W., Schuster, P. (1987) *Biophys.Chem.* **26**: 123

[17] Fontana, W., Schnabl, W., Schuster, P. (1989) *Phys.Rev.A* **40**: 3301

[18] Schuster, P. (1991) *Complex optimization in an aritificial RNA world.* In: Farmer, D., Langton, C., Rasmussen, S., Taylor, C., eds. Artificial Life II. SFI Studies in the Sciences of Complexity, Vol.XII. Addison-Wesely Longman, Redwood City (Cal.)

[19] Eigen, M., McCaskill J., Schuster, P. (1988) *J.phys.Chem.* **92**: 6881

[20] Schuster, P., Swetina, J. (1988) *Bull.Math.Biol.* **50**: 635

[21] Eigen, M., McCaskill J., Schuster, P. (1989) *Adv.Chem.Phys.* **75**: 149

[22] Biebricher, C.K., Eigen, M., Gardiner, jr., W.A. (1991) *Quantitative analysis of selection and mutation in self-replicating RNA.* In: Peliti, L., ed. Biologically Inspired Physics. NATO Advanced Study Series, in press.

[23] Jaeger, J.A., Turner, D.H., Zuker, M. (1989) *Proc.Natl.Acad.Sci.USA* **86**: 7706

[24] Fontana, W., Konings, D.A.M., Schuster, P. (1991) *Statistics of RNA secondary structures.* Preprint.

[25] Sankoff, D., Morin, A.-M., Cedergren, R.J. (1978) *Can.J.Biochem.* **56** : 440

[26] Cech, T.R. (1988) *Gene* **73** : 259

[27] Le, S.-Y., Zuker, M. (1990) *J.Mol.Biol.* **216** : 729

[28] Zuker, M., Sankoff, D. (1984) *Bull.Math.Biol.* **46** : 591

[29] Zuker, M. (1989) *Science* **244** : 48

[30] Jaeger, J.A., Turner, D.H., Zuker, M. (1990) *Methods in Enzymology* **183** : 281

[31] McCaskill, J.S. (1990) *Biopolymers* **29** : 1105

[32] Hamming, R.W. (1986) *Coding and Information Theory.* 2nd Ed. Prentice Hall. Englewood Cliffs (N.J.), pp.44-47.

[33] Maynard Smith, J. (1970) *Nature* **225** : 563

[34] Shapiro, B.A. (1988) *CABIOS* **4** : 387

[35] Shapiro, B.A., Zhang, K. (1990) *Comparing multiple RNA secondary structures using tree comparisons. CABIOS*, in press.

[36] Sankoff, D., Kruskal, J.B. (1983) *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison.* Addison-Wesley. Reading (Mas.).

[37] Tai, K.-C. (1979) *J.Ass.Computing Machinery* **26** : 422

[38] Hogeweg, P., Hesper, B. (1984) *Nucleic Acids Research* **12** : 67

[39] Konings, D.A.M. (1989) *Pattern analysis of RNA secondary structure.* Proefschrift, Rijksuniversiteit te Utrecht.

[40] Konings, D.A.M., Hogeweg, P. (1989) *J.Mol.Biol.* **207** : 597

[41] Fontana, W., Konings, D.A.M., Stadler, P.F., Schuster, P. (1991) *Quantitative comparison and statistics of RNA secondary structures.* Preprint.

[42] Karlin, S., Taylor, H.M. (1975) *A first course in stochastic processes.* 2nd Ed. Academic Press. New York, pp.455-461.

[43] Weinberger, E.D. (1990) *Biol.Cybern.* **63** : 325

[44] Sherrington, D., Kirkpatrick, S. (1975) *Phys.Rev.Letters* **35** : 1792

[45] Stadler, P.F., Schnabl, W. (1991) *The landscape of the traveling salesman problem.* Preprint.

[46] Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T., Turner, D.H. (1986) *Proc.Natl.Acad.Sci.USA* **83** : 9373

[47] Fontana, W., Stadler, P.F., Griesmacher, T., Weinberger, E.D., Schuster, P. (1991) *Statistical properties of RNA free energy landscapes. A study by random walk techniques.* Preprint.

Contents