Statistics of RNA Secondary Structures

Walter Fontana
Danielle A. M. Konings
Peter F. Stadler
Peter Schuster

SFI WORKING PAPER: 1992-02-007

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu
Statistics of RNA Secondary Structures †

By

WALTER FONTANA \textsuperscript{a,b,c}, DANIELLE A. M. KONINGS \textsuperscript{d}, PETER F. STADLER \textsuperscript{a}
and PETER SCHUSTER \textsuperscript{a,c,e,*}

\textsuperscript{a} Institut für Theoretische Chemie der Universität Wien

\textsuperscript{b} Theoretical Division, Los Alamos National Laboratory

\textsuperscript{c} Santa Fe Institute

\textsuperscript{d} Department of Molecular, Cellular and Developmental Biology, University of Colorado at Boulder

\textsuperscript{e} Institut für Molekulare Biotechnologie, Jena

Submitted to Biopolymers

† Dedicated to Professor Manfred Eigen

* 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 pks@imb.uni-jena.dbp.de
Contents

1. Introduction 2
2. Statistics of elements of RNA secondary structures 4
3. Tree representations and tree distances 10
4. Complex combinatorial maps, landscapes, and density surfaces 14
5. Autocorrelation functions and correlation lengths 19
6. Comparison with natural sequences 21
   Acknowledgements 24
References 26
1. Introduction

A great variety of biopolymers were studied extensively by sequence analysis, X-ray diffraction and spectroscopic techniques. Their molecular structures are well known by now. Despite the availability of very detailed information on many individual biomolecules very little, if anything, is known about the statistics of structural features in large ensembles of biopolymers, as well as about the stability of structures against modifications in the sequences. The difficulty in obtaining this information is twofold:

- the numbers of molecules studied so far are still too small to allow for significant statistical analysis, and
- the probes taken are anything but random since they represent the present-day outcome of a long evolutionary selection process.

A possible strategy to overcome this fundamental lack of knowledge is to compute large statistical ensembles of biopolymer structures. At this point, however, one is facing another serious problem of current theoretical biophysics: computation of spatial structures of biopolymers in three dimensions is highly time consuming and unreliable in general. In the case of the two so far best studied classes of biopolymers, proteins, DNA and RNA molecules, the predictive power of the available algorithms is poor.

Secondary structures of RNA molecules are much easier to predict. They are mainly determined by the conventional base pairing rules of RNA: \( G=\text{C} \), \( A=\text{U} \) and \( G-\text{U} \). In this paper we also consider the non-natural xanthine–2,6-diaminopyrimidine pair, \( \text{X}=\text{K} \), which was recently incorporated enzymatically into synthetic RNA and DNA molecules\(^1\). Base pairing and base pair stacking energies are generally larger than those of other interactions involved in the formation of spatial structures and hence, it is meaningful to partition the folding of the primary structure of an RNA molecule into the three-dimensional tertiary structure into two steps:
Abstract.

Large ensembles of RNA sequences are folded into secondary structures with minimum free energies. Four nucleotide *al*phabets are used: two binary alphabets, AU and GC, the biophysical AUGC and the synthetic GCXK alphabet. They define base pairing rules, and by their physical nature also the strengths of the base pair interactions. All quantities presented here depend strongly on the particular alphabet chosen. RNA secondary structures are partitioned into structural elements, such as stacks, loops, joints and free ends. Statistical properties of these elements are computed for different chain lengths up to $\nu = 100$. The results obtained from the statistics of random ensembles are compared with the data derived from natural RNA molecules with similar base frequencies.

Secondary structures are represented as trees. A quantitative measure for the distance between two structures, the *tree distance* $d_t$, is obtained by means of tree editing. Two different, but formally equivalent tree representations are introduced and compared in actual computations of RNA structures.

We introduce a structure density surface as the the conditional probability $P(t|h)$ of two structures having tree distance ($d_t = t$) given that the sequences that fold into them have Hamming distance ($d_h = h$). Structure density surfaces provide insight into the *shape space* of RNA secondary structures. Nearly the entire range of tree distances is covered with considerable probability already at small Hamming distances from a typical sequence. This suggests that the vast majority of possible structures occur within a fairly small neighbourhood of any random sequence. Correlation lengths for secondary structures in their tree representations are computed from probability densities. They are appropriate measures of the complexity or *ruggedness* of structure landscapes.

Keywords.

Complex combinatorial landscapes -- Correlation length -- RNA secondary structures -- Shape space covering -- Random RNA sequences -- Tree editing
(1) folding of the primary structure, a string of bases, into a two-dimensional, planar secondary structure by base pair formation, and

(2) formation of the spatial structure by folding the planar secondary structure into a three-dimensional object.

The partitioning of the folding process into two steps is not free from arbitrariness. Pseudoknots are commonly considered as elements of the tertiary structure, but, in principle, they could be also incorporated into secondary structures. Other additional base pairs which are not compatible with a planar structure are attributed to tertiary structures. Both steps, (1) and (2), follow a minimum free energy criterion unless suboptimal folding patterns are to be determined.

Another problem turns out to be prohibitive for the handling of large ensembles of tertiary structures. A three-dimensional structure can be stored only by listing Cartesian coordinates of many thousands of atoms. Secondary structures, on the other hand, can be stored in compressed form. As we show in section 2 the encoded structures do not need more storage capacity than the strings representing primary sequences. Still, processing of millions of secondary structures provides some memory problems which, however, can be overcome by the present computation facilities.

The presently predominantly used folding algorithms for RNA assume that secondary structures are partitioned into elements defined in section 2 which contribute additively to the free energies of the molecules. In reality the non-additive contributions are fairly small, and hence are in good approximation attributed to the tertiary structure. The algorithms are derived from a method based on dynamic programming which was originally conceived by Zuker, Stiegler and Sankoff², ³. It was primarily designed to compute the minimum free energy structure, but derivative algorithms allow to obtain suboptimal foldings as well⁴–⁶. Alternatively one may consider suboptimal foldings with the corresponding Boltzmann weights and compute partition functions for RNA secondary structures di-
rectly\textsuperscript{7}. The empirical parameters used in the folding algorithm were updated and summarized some years ago\textsuperscript{8}.

In this paper we shall be dealing exclusively with minimum free energy secondary structures computed by a derivative of the Zuker algorithm. Our computer code was originally designed for fast folding as part of a simulation package for molecular evolution\textsuperscript{9,10}. In the present version of the software package, which includes a statistics program as well as tree editing routines\textsuperscript{11}, a currently updated version of the empirical parameter set was used\textsuperscript{12}. For the \textbf{XK} base pair we use the \textbf{GC} parameter set which seem to come closest to the base pairing strength of the synthetic base pair\textsuperscript{1}.

2. Statistics of elements of RNA secondary structures

The folding algorithm is a procedure which converts an RNA primary sequence, say

$\mathbf{I}_k = \{\text{AUGCUUGGACGUGCAGUCCAGUCAG...AAACGC}\}$,

into a secondary structure $S_k = S(\mathbf{I}_k)$ where $S(.)$ stands for the folding algorithm which computes a unique structure for every sequence $\mathbf{I}_k$. An example is shown in figure 1. Many sequences, however, may fold into the same secondary structure. This fact makes the reverse folding problem – the problem to determine all sequences which fold into a given secondary structure – a particularly hard task.

\textbf{Fig. 1:} An example for folding an RNA sequence $\mathbf{I}_k$ into a secondary structure $S_k$ and its conversion into a (full) tree $T_k$. In this tree representation single stranded bases are shown as open circles (o) and base pairs as a full circles (•), respectively. A root (●), not corresponding to a physical unit of the RNA, is added. The full tree $T_k$ is transformed into a homeomorphically irreducible
tree (HIT) $H_k$ by assigning a weight $w$ to every node of the HIT which counts the number of nodes which are contracted into a single one.

**Fig. 2:** Conventional structure elements of RNA secondary structures. The elements are denoted by S for stack, H for hairpin loop, B for buldge, I for internal loop, M for multiloop, J for joint and E for free end. Individual nucleotides are indicated by •.

RNA secondary structures $S_k$ are strictly planar graphs. Planarity essentially means that unpaired bases inside a loop are not allowed to pair with unpaired bases outside of that loop. A secondary structure is viewed conventionally as a combination of structure elements which fall into seven classes (Figure 2):

1. **stacks** (S) which are double helical regions consisting of stacked base pairs,
2. **hairpin loops** (H) representing stretches of unpaired bases which close terminal stacks,
3. **bulges** (B) which connect two stacks by an unpaired stretch,
4. **internal loops** (I) joining two stacks with two single stranded stretches,
5. **multiloops** (M) consisting of several single stranded stretches which connect more than two stacks,
6. **joints** (J), which are stretches of unpaired bases joining freely movable substructures, and
7. **free ends** (E).

Nucleotides in joints and free ends are often termed external bases. Isolated single base pairs are considered as stacks as well. The degree of a loop is the number of stacks connected to it. It is often useful to lump loops of all degrees together into one class and to consider, for example, the total number of loops

$$n_L = n_H + n_B + n_I + n_M,$$

which must be identical to the number of stacks, $n_L = n_S$. 


We report on statistical properties of secondary structures computed for different chain lengths $\nu$ and for different base pairing alphabets, the biophysical alphabet $\text{AUGC}$ and the synthetic $\text{GCXK}$ alphabet (both having $\kappa = 4$ types of digits), and the two binary alphabets, $\text{AU}$ and $\text{GC}$ ($\kappa = 2$). Base frequencies were chosen around the most probable distributions, $(0.25,0.25,0.25,0.25)$ or $(0.5,0.5)$, respectively. Such base frequencies may be obtained in actual computations of large ensembles simply by the assumption of equal probabilities for all point mutations. The four alphabets are chosen for obvious reasons: $\text{AU} \leftarrow \text{AUGC} \leftarrow \text{GC}$ spans the entire region of natural RNA molecules and the three natural alphabets represent the extreme cases. Analogous studies for intermediate base compositions are under way. The synthetic $\text{GCXK}$ alphabet is interesting since it allows to study the properties of a four letter alphabet with two complementary base pairs without the complications of different base pair strengths and additional non-standard ($\text{G–U}$) interactions as in the biophysical set. A few hundred thousands of random RNA sequences were folded and the secondary structures were analyzed with respect to frequency of occurrence and size of the various structural elements. Unstable structures, these are structures with free energies $f \geq 0$, are not considered for structure statistics. The distribution of free energies and other features related to thermodynamical stabilities of secondary structures are discussed elsewhere.

**Fig. 3:** The mean number of base pairs ($n_{\text{BP}}$) as a function of the chain length $\nu$. Values are shown for binary $\text{GC}$-sequences (○), for binary $\text{AU}$-sequences (○), for four letter $\text{GCXK}$-sequences with $\text{GC}$ parameters (*), and for natural $\text{AUGC}$-sequences (●).

The mean number of base pairs in secondary structures ($n_{\text{BP}}$) increases linearly with the chain length $\nu$ for sufficiently long sequences (Fig.3). Deviations at small chain lengths ($\nu < 50$) are found with $\text{AU}$-, $\text{AUGC}$-, and $\text{GCXK}$-sequences.
The influence of the base pairing alphabet is interpreted in straightforward manner by considering the *stickiness* $P$ of the sequences which is understood as the probability that two arbitrarily chosen bases can form a base pair. Let $p_i$ be the frequency of digit "i" which is given by $n_i/\nu$ with $n_i$ being number of digits of type "i" in the sequence. Clearly we have $\sum_{i=1}^{n} p_i = 1$, and we obtain for the stickiness in the four alphabets:

\begin{align}
P_{\text{AU}} &= 2p_{\text{AU}} \quad \text{and} \quad P_{\text{GC}} = 2p_{\text{GC}} , \quad (2a) \\
P_{\text{AUGC}} &= 2\left(p_{\text{AU}} + p_{\text{UPG}} + p_{\text{GPC}}\right) , \quad (2b) \\
P_{\text{GCXK}} &= 2\left(p_{\text{GPC}} + p_{\text{XPK}}\right) . \quad (2c)
\end{align}

For the base compositions used here we find $P_{\text{AU}} = P_{\text{GC}} = 0.5$, $P_{\text{AUGC}} = 0.375$, and $P_{\text{GCXK}} = 0.25$. As expected, and as seen in figure 3, the pure GC-sequences are leading with respect in the number of base pairs: they have the highest possible stickiness and form the strongest base pairs. Sterical constraints, for example those in loops, are more easily compensated by GC pairs than by AU pairs. Hence AU-sequences form fewer base pairs on the average than GC-sequences. Sequences derived from four letter alphabets are less sticky and form still fewer base pairs. In addition the slope in the $(n_{BP}/\nu)$-plot is smaller too. The fact that AUGC- and GCXK-sequences have almost the same mean numbers of base pairs is fortuitous: the former are more sticky, the latter form stronger base pairs and the two effects cancel by accident.

**Fig. 4:** The mean number of loops $(n_L)$ or stacks $(n_S)$ as a function of the chain length $\nu$. Values are shown for binary GC-sequences ($\circ$), for binary AU-sequences ($\diamond$), for four letter GCXK-sequences with GC parameters ($\ast$), and for natural AUGC-sequences ($\bullet$).

Another quantity that is useful to characterize secondary structures is the mean number of loops $n_L$ per structure. Since every loop is closed by exactly one
stack the mean number of stacks is identical to the mean number of loops \( n_S = n_L \).

As shown in figure 4 it increases also linearly with chain length \( \nu \). Sequences with lower stickiness values have on the average more loops than stickier sequences.

The effect of base pair strength is even more pronounced than that of stickiness: weak base pairing results in fewer stacks and hence the structures derived from **AUCG-** or **AU-**sequences have fewer loops than their **GCXK** or **GC** counterparts.

**Fig. 5:** The mean number of components \( n_C \) connected by \( n_J = n_C - 1 \) joints as a function of the chain length \( \nu \). Values are shown for binary **GC-**sequences (○), for binary **AU-**sequences (○), for four letter **GCXK-**sequences with **GC** parameters (★), and for natural **AUGC-**sequences (●).

A secondary structure consists of one, two or more components which are connected by joints. The mean number of components \( n_C \) shows a characteristic lag phase before it starts to increase with increasing chain length \( \nu \). This lag phase reflects the fact that a certain minimum chain length is required in order to form structures with two or more components. The lag phase is more pronounced in structures built from alphabets with weaker base pairs (**AUGC, AU**). The increase of \( n_C \) with \( \nu \) is much stronger in the case of the four letter alphabets. The data shown in figure 5 suggest that this increase is roughly linear. In order to be able to study large ensembles of longer sequences the folding algorithm was adapted to a parallel computer\(^{15}\). These computations have shown, however, that the chain length dependence of the number of components is more complicated: it seems to be either logarithmic, or \( n_C \) turns eventually into an asymptotic linear increase at chain length substantially larger than \( \nu = 1000 \).

**Fig. 6:** The mean degree of loops \( (n_{LD}) \) as a function of the chain length \( \nu \). Values are shown for binary **GC-**sequences (○), for binary
AU-sequences (●), for four letter GCXK-sequences with GC parameters (●), and for natural AUGC-sequences (●).

The average degree of loops (Fig.6) is in the range $1 < n_{LD} < 2$. It converges to a constant value with increasing chain length $\nu$. Structures derived from sequences with strong base pairs (GC, GCXK) have more higher branches than those obtained from AUGC- and AU-sequences.

**Fig. 7:** The mean number of base pairs in one stack ($n_{st}$) as a function of the chain length $\nu$. Values are shown for binary GC-sequences (○), for binary AU-sequences (●), for four letter GCXK-sequences with GC parameters (●), and for natural AUGC-sequences (●).

**Fig. 8:** The mean number of bases in one loop ($n_{lp}$) as a function of the chain length $\nu$. Values are shown for binary GC-sequences (○), for binary AU-sequences (●), for four letter GCXK-sequences with GC parameters (●), and for natural AUGC-sequences (●).

Mean stack sizes ($n_{st}$) and mean loop sizes ($n_{lp}$) converge to almost constant values at fairly small chain lengths ($\nu \approx 50$) already as shown in the figures 7 and 8. Apparently the convergence of $n_{st}$ and $n_{lp}$ in structures from GC-sequences is slower than in structures from the other three alphabets. Stickiness is important for stack sizes: on the average the two letter sequences form longer stacks than AUGC-sequences. The mean stack lengths of GCXK-sequences are shortest. Weak base pairing makes nucleation of stacks more difficult. Stacks are, therefore, longer on the average, as can be seen from a comparison of AU- and GC-sequences. Large loops are favoured by both low stickiness and weak base pairs since a weak closing pair is more likely to be destabilized by loop strain than a strong one. Thus GC-sequences form the smallest loops on average and in long sequences $n_{lp}$ has approximately the same value as the smallest allowed hairpin loop, $n_{lp} = 3$. 
So far only mean values of the statistical properties of RNA secondary structures were considered. Resolved probability densities for stack sizes, loop sizes and loop degrees will be presented in section 6 where the data derived from random sequences will be compared with those obtained from natural sources with similar base compositions.

**Fig. 9:** The mean number of unpaired external bases \( (n_{\text{ext}}) \) as a function of the chain length \( \nu \). Values are shown for binary GC-sequences (○), for binary AU-sequences (●), for four letter GCXK-sequences with GC parameters (⋆), and for natural AUGC-sequences (●).

The last quantity to be considered here is the mean number of external bases \( (n_{\text{ext}}) \). Its increase with the chain length \( \nu \) appears to be more complex than linear. This is no surprise since the number of external bases contains contributions from terminal free bases and free bases in joints. The number of joints in turn shows a complex increase with the chain length (Fig.5). The strength of base pairs has a strong influence on \( n_{\text{ext}} \): weak pairing implies larger mean numbers of external digits. Interestingly \( n_{\text{ext}} \) is almost constant for GC-sequences where the number of components converges to a constant value for longer sequences \( (\nu > 300) \) too\(^{15}\).

3. **Tree representations and tree distances**

RNA secondary structures are often represented by trees\(^3,16,17\). The so-called *mountain representation* is an alternative graphical method for the comparison of RNA secondary structures\(^{18-20}\). A secondary structure \( S_k \) is converted into a tree \( T_k \) by application of a few rules (Fig.1). The conversion starts with a root which does not correspond to a physical unit of the RNA molecule. It is introduced to yield a single tree representation by preventing the formation of a
tree forest for RNA molecules with free ends. The topmost left branch of the tree corresponds to the 5'-end of the polynucleotide chain. Every single stranded base of the secondary structure is represented by a leaf node, every base pair by an internal node, and all neighbourhood relations are preserved. The tree is equivalent to the secondary structure since no information is lost when the structure is converted into the tree and vice versa.

As shown in figure 1 the trees $T_k$ can be rewritten as homeomorphically irreducible trees (HITs) which will be denoted by $H_k$. The apparently simpler tree structure of the HIT is compensated by the assignment of weights ($w$) to the internal nodes and leaves. A weight reflects the number of nodes or leaves in the full tree $T_k$ which are lumped into a single node or leaf of the HIT representation. Strictly speaking, the HIT proposed here is homeomorphical only with respect to the structure given by the internal nodes. Contraction of two or more leaves into one weighted leaf changes the branching structure. Nevertheless the transformation from the full tree to the HIT retains complete information on the structure. Secondary structure, full tree as well as HIT are, therefore, equivalent:

$$I_k \Rightarrow S_k \iff T_k \iff H_k .$$ (3)

Tree editing provides a distance in the space of trees, and hence in the space of RNA secondary structures. A tree is transformed into another tree by a series of tree-editing operations for which predefined costs are accounted $^{16, 17, 21, 22}$. The distance between two trees,

$$d(T_j, T_k) = d_t(j, k) ,$$ (4)

is given by the sum of the costs along a editing path which is minimal with respect to this sum. It can be shown that this tree distance forms a metric in the space of trees $^{22}$. The parameters used in tree editing are summarized in table 1. The editing operations preserve the relative traversal order of the tree nodes. Tree
**Table 1**: Cost table for tree edit operations in the full tree representation and in the homeomorphically irreducible tree representation (HIT) and the edit cost parameters used in this paper. Weights in the HIT representation are denoted by $w$ and $v$.

<table>
<thead>
<tr>
<th>Edit operation</th>
<th>Symbolic notation</th>
<th>Edit parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Symbol</td>
<td>Cost</td>
</tr>
<tr>
<td><strong>Full Tree Edit Operations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single base insertion</td>
<td>$0 \rightarrow o$</td>
<td>$\delta_o$</td>
</tr>
<tr>
<td>Base pair insertion</td>
<td>$0 \rightarrow \bullet$</td>
<td>$\delta_*$</td>
</tr>
<tr>
<td>Single base deletion</td>
<td>$o \rightarrow 0$</td>
<td>$\delta_o$</td>
</tr>
<tr>
<td>Base pair deletion</td>
<td>$\bullet \rightarrow 0$</td>
<td>$\delta_*$</td>
</tr>
<tr>
<td>Relabel</td>
<td></td>
<td>$\rho$</td>
</tr>
<tr>
<td>Substitution</td>
<td>$\bullet \leftrightarrow o$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td><strong>HIT Edit Operations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insertion of single bases</td>
<td>$0 \rightarrow o w$</td>
<td>$w \cdot \delta_o$</td>
</tr>
<tr>
<td>Insertion of base pairs</td>
<td>$0 \rightarrow \bullet w$</td>
<td>$w \cdot \delta_*$</td>
</tr>
<tr>
<td>Deletion of single bases</td>
<td>$w o \rightarrow 0$</td>
<td>$w \cdot \delta_o$</td>
</tr>
<tr>
<td>Deletion of base pairs</td>
<td>$w \bullet \rightarrow 0$</td>
<td>$w \cdot \delta_*$</td>
</tr>
<tr>
<td>Relabel of single bases</td>
<td>$w o \leftrightarrow o v$</td>
<td>$</td>
</tr>
<tr>
<td>Relabel of base pairs</td>
<td>$w \bullet \leftrightarrow \bullet v$</td>
<td>$</td>
</tr>
<tr>
<td>Substitution</td>
<td>$w \bullet \leftrightarrow o v$</td>
<td>$\sigma \min{w,v} +</td>
</tr>
</tbody>
</table>

Editing can therefore be viewed as a generalization of sequence alignment. In fact, for trees that consist solely of leaves, tree editing becomes the standard sequence alignment.

Using the parameters of table 1 for editing operations on weighted trees,
distances between HITs can be computed as well:

$$d(H_j, H_k) = d^{(HIT)}(j, k).$$

(5)

Tree distances between full trees and HITs fulfil the general relation

$$d_t \leq d_t^{(HIT)}.$$  

(6)

A proof will be given in a forthcoming paper\textsuperscript{23}.

Alternatively, distances between secondary structures can be computed by direct end-to-end alignment of encoded secondary structures as discussed by Paulien Hogeweg and coworkers\textsuperscript{18–20} (in case of equal length encodings this becomes a Hamming distance). Later on we shall see that several important properties do not depend on the distance measure applied. In other words, we obtain properties that are generic with respect to structure distances computed as Hamming distances, HIT distances $d_t^{(HIT)}$ or distances between full trees $d_t$.

Tree representations in full resolution make it often difficult to focus on the major structural features of RNA molecules since they are often overloaded with irrelevant details. Coarse-grained tree representations were invented to solve this problem\textsuperscript{16,17}. The HITs introduced here are particularly well suited for coarse-graining because the weights $w$ can be used as resolution parameters. Coarse-graining of HITs will be the subject of a forthcoming publication\textsuperscript{23}.
4. Complex combinatorial maps, landscapes, and density surfaces

Comparison of RNA sequences is a common tool in molecular evolution. In the simplest case the distance between two sequences of the same length $\nu$ is given by the Hamming metric \( d_h(k, \ell) \) which counts the number of different digits in the two end-to-end aligned sequences $I_k$ and $I_\ell$. This is the distance measure we shall be concerned with here. The set of all sequences of a given length $\nu$ is of combinatorial complexity: an alphabet with $\kappa$ letters yields $\kappa^\nu$ different sequences. Embedding this set in euclidean space such that pairs of sequences with Hamming distance $d_h = 1$ are closest neighbours yields the so-called sequence space \(^{24,26}\) for which $d_h$ is a metric.\(^ {25}\) The sequence space of binary sequences ($\kappa = 2$) is a hypercube of dimension $\nu$. In the case of four letter alphabets ($\kappa = 4$) the sequence space is a more complex object which can be understood as a union of two hypercubes with the appropriate connections. Another variant of sequence space is the protein space introduced by John Maynard-Smith.\(^ {27}\)

Secondary structures $S_k$ or trees $T_k$, respectively, form also a space of combinatorial complexity with the tree distance $d_t(k, \ell)$ being a metric on it. The notion of a shape space, originally conceived for antibody-antigen recognition in theoretical immunology\(^ {28}\), appears to be an appropriate characterization for RNA secondary structures as well.

It is useful to consider RNA folding in the formal concepts of sequence and shape spaces. The process of folding assigns a secondary structure $S_k$, or a tree $T_k$ respectively, to an element $I_k$ of the sequence space. It can be understood therefore as a mapping $\Phi$ from a sequence space $X$ into a shape space $Y$:

$$\Phi : (X, d_h) \implies (Y, d_t). \quad (7)$$

The elements of both spaces are discrete structures of combinatorial complexity. Both spaces are endowed with appropriate metrics: Hamming distance $d_h$, and
tree distance $d_t$. Accordingly, a mapping of this kind is called a complex combinatorial map (CCM). If the property to be assigned to the sequence is a scalar quantity like the free energy, the corresponding mapping into the real numbers,

\[ \mathcal{L} : (X, d_h) \rightarrow (\mathbb{R}^1), \quad (8) \]

is a complex combinatorial landscape (CCL). Landscapes of this class are generalizations of mappings from genotype space into fitness values which were introduced as fitness landscapes into biology already by Sewall Wright. More recently fitness landscapes saw a revival as rugged landscapes on which evolutionary adaptive walks are assumed to take place. The concept of complex combinatorial landscapes turned out to be a useful heuristic principle for the understanding of adaptation in evolution and coevolution. When it comes to do actual computations or if one wants to use these landscapes for visual comparisons, a direct representation of CCLs is rather inefficient and often too sophisticated.

A two dimensional probability density $P(., h)$ is proposed for gaining insight into the details of RNA folding without dealing with CCMs explicitly. A tree distance density surface, for example, expresses the joint probability of two RNA sequences of chain length $\nu$, $I_i$ and $I_j$, having Hamming distance $d(I_i, I_j) = d_h(i, j) = h$, as well as having a tree distance $d(T_i, T_j) = d_t(i, j) = t$ between their secondary structures. The surface $P(t, h)$ is biased along the Hamming distance ($h$) axis with the distribution

\[ p(h) = \binom{\nu}{h} (\kappa - 1)^h \kappa^{-\nu}. \quad (9) \]

The bias reflects combinatorics: there are $\kappa^\nu \cdot p(h)$ sequences with Hamming distance $h$ from a given sequence. In case we are dealing with binary sequences $p(h)$ is simply the binomial distribution. In order to compensate for the bias we consider the conditional probability of finding a tree distance $t$ between two structures whose sequences have a Hamming distance $h$:

\[ P(t|h) = \frac{P(t, h)}{p(h)} \approx \frac{n(t, h)}{\sum_{l=0}^{\text{max}} n(t, h)}. \quad (10) \]
As indicated in equation (10) the conditional probability $P(t|h)$ is computed by sampling tree distances of pairs of sequences with Hamming distance $h$. By $n(t, h)$ we denote the number of pairs of sequences in the sample which have Hamming distance $h$ and tree distance $t$; $t_{\text{max}}$ is used for the maximal tree distance. The total sample size is $N = \sum_{t=0}^{t_{\text{max}}} \sum_{h=0}^\nu n(t, h)$ and from the definition of conditional probabilities follows $\sum_{t=0}^{t_{\text{max}}} P(t|h) = 1$ for each $h$.

A sampling technique is applied here which turned out to be very convenient for density surfaces since it aims directly at a computation of conditional probabilities. It may be characterized as uniform sample statistics and proceeds as follows:

(i) we choose a reference sequence at random,
(ii) we sample exactly $\ell$ sequences at each Hamming distance $h = 1, 2, \ldots, \nu - 1$ from the reference sequence,
(iii) we fold these sequences into secondary structures and compute tree distances relative to structure of the reference sequence,
(iv) we sample them in $(t, h)$ boxes just be counting numbers of instances, and
(v) continue at item (i) until convergence or the desired accuracy, respectively, has been achieved after computation of say $r$ reference points.

By definition we have in this case

$$\sum_{t=0}^{t_{\text{max}}} n(t, h) = r \cdot \ell = \text{const.} \quad (11)$$

and thus the sampling procedure corrects directly for the bias of the binomial distribution. The sample size now is $N = r \cdot \ell \cdot (\nu - 1)$.

**Fig. 10:** Probability density surfaces for tree distances of secondary structures of AUGC- and GC-sequences of chain length $\nu = 100$.

Two examples of tree distance density surfaces are shown in figure 10. Both surfaces show an overall shape and, superimposed upon it, rich and bizarre looking
details. Density surfaces of binary sequences exhibit a kind of symmetry. This symmetry is interesting since it can be only approximate for two reasons:

(1) the complementary sequence of the reference sequence is obtained by base complementation and swapping 3'- and 5'-end, and, in general, is not the sequence at Hamming distance $\nu$, and

(2) complementary sequences do not have the same free energies and need not have identical secondary structures since the terminal mismatches at the end of stacks introduce asymmetries.

Tree distance density surfaces for binary sequences show three major maxima: one at $h=0$, one at $h=\nu$ and one at $h=\nu/2$. The first one trivially appears at $t=0$ and the second one is either at or very close to $t=0$. The central maximum of the tree distance density surface lies rather far away from $t=0$ and thus indicates that the secondary structures of a randomly chosen pair of sequences – which most likely has a Hamming distance $h=\nu/2$ – are substantially different.

As expected tree density surfaces computed from sequences of four letter alphabets, for example from AUGC-sequences (Fig.10), lack the approximate symmetry found with binary sequences. The first half of the density surface ($0 \leq h \leq \nu/2$) resembles closely the corresponding part of the surface of binary sequences. The second half ($\nu/2 \leq h \leq \nu$), however, is like a mountain ridge parallel to the line $t=0$ and indicates that the distribution of tree distances is more or less independent of the Hamming distance $h$ at large distances from the reference sequence.

**Fig. 11:** Probability density surfaces for tree distances of full trees (A) and HITs (B) from structures of GC-sequences of chain length $\nu=120$.

The overall shape of the tree density surfaces is not restricted to full trees. The same features are observed with density surfaces derived from HITs (Fig.11) or from other structure distance measures obtained by direct alignment of the
encoded secondary structures\textsuperscript{35}. These results demonstrate that the overall shape of tree density surfaces is a robust statistical feature of RNA secondary structures.

Inspection of the tree density surface in the neighborhood of $h=0$ shows that large values of tree distances – coming close to the maximum – are common at small Hamming distances already. Since this general aspect of density surfaces for RNA secondary structures does neither depend on the chain length $\nu$, nor on the base pairing alphabet, nor on the particular measure for structure distances, we turn it into a conjecture on RNA shape space covering: almost all typical RNA secondary structures, these are structures obtained from random sequences, occur already within a small neighbourhood of any typical sequence.

A free energy density surface is constructed and computed in complete analogy if the tree distance $d_t(i,j) = t$ is replaced by the absolute value of the difference in free energies,

$$d_f(i,j) = |f(S_i) - f(S_j)| = f.$$  

As examples we show energy density surfaces for AUGC- and GC-sequences of chain length $\nu = 100$ (Fig.12). Free energy density surfaces differ largely from tree distance density surfaces in their overall shapes. Again we see rich and very rugged superimposed details. In slices parallel to the $f$-axis the general shape is described roughly by the positive half of a Gaussian curve. This indicates that the distribution of free energies in sequence space is approximately normal – negative values of $f$ are excluded by the definition in equation (12). The distribution and other properties related to free energies of RNA secondary structures are discussed elsewhere\textsuperscript{14}. As with tree distances free energy density surfaces derived from GC-sequences show an approximate symmetry which is lacking in those derived from AUGC-sequences.

**Fig. 12:** Probability density surfaces for the free energies of secondary structures of AUGC- and GC-sequences of chain length $\nu=100$. 
5. Autocorrelation functions and correlation lengths

Probability density surfaces are not only a means of illustration, they can also be used to compute statistical quantities of CCMs or CCLs. As an example we consider the autocorrelation functions of tree distances and free energies which can be expressed by means of mean square distances:

\[
\varrho(h) = 1 - \frac{\langle (X_i - X_{i+h})^2 \rangle}{\langle (X_i - X_j)^2 \rangle} = 1 - \frac{\langle d^2(h) \rangle}{\langle d^2 \rangle}.
\]

The expression \( \langle d^2 \rangle \) is the mean square distance - tree distance \( d_t \) or free energy difference \( d_f \) - measured over the entire sequence space. The mean square distance derived from sequences with given Hamming distance \( h \) is denoted by

\[
\langle d^2(h) \rangle := \langle d^2(i, h | d_h(i, h) = h) \rangle.
\]

Both mean square distances can be computed from the density surface \( P(t|h) \). Let us consider tree distances first. The conditional mean square distance is simply the expectation value of \( t^2 \) computed for a given Hamming distance \( h \)

\[
\langle d_t^2(h) \rangle = \sum_{t=0}^{t_{\text{max}}} t^2 P(t|h) \approx \frac{\sum_{t=0}^{t_{\text{max}}} t^2 n(t, h)}{\sum_{t=0}^{t_{\text{max}}} n(t, h)}.
\]

Recalling that the mean square distance on the entire sequence space can be expressed as a weighted sum of the conditional mean square distances \( \langle d_t^2(h) \rangle \) we find

\[
\langle d_t^2 \rangle = \sum_{h=0}^{\nu} \langle d_t^2(h) \rangle \cdot p(h) \approx \sum_{h=0}^{\nu} \frac{\sum_{t=0}^{t_{\text{max}}} t^2 n(t, h) p(h)}{\sum_{t=0}^{t_{\text{max}}} n(i, h)}.
\]

Rewriting equation (14) yields the autocorrelation function in terms of the sampling array \( n(t, h) \),

\[
\varrho_t(h) = 1 - \frac{\sum_{t=0}^{t_{\text{max}}} t^2 n(t, h)}{\sum_{h=0}^{\nu} \sum_{t=0}^{t_{\text{max}}} t^2 n(t, h) p(h)},
\]

which is applicable to numerical evaluation if the sample size is the same in each mutant class - uniform sampling statistics, as discussed in equation (11).
Fig. 13: Correlation lengths of tree distances ($\ell_t$) of RNA molecules in their most stable secondary structures as functions of the chain length $\nu$. Values are shown for binary GC-sequences (o), for binary AU-sequences (o), for four letter GCXK-sequences with GC parameters (*), and for natural AUGC-sequences (●).

Autocorrelation functions of tree distances $g_t(h)$ are used to compute correlation lengths of RNA trees $\ell_t$ by an empirical procedure: the point $\ln g_t(\ell_t) = -1$ is evaluated from an $\ln g_t(h)$ plot by means of a least root mean square deviation fit. The tree correlation length is a useful measure for the stability of RNA secondary structures against mutation. As we conclude from figure 13 the correlation length increases roughly linearly with the chain length $\nu$ and depends strongly on the base pairing alphabet. Binary sequences, AU or GC, are much more likely to change their minimum free energy structures on point mutations than GCXK-sequences. Natural AUGC-sequences are still less sensitive to point mutations. This is apparently a consequence of the possibility to form G–U pairs in stacks which makes more changes in the sequences tolerable in secondary structures.

Autocorrelation functions can be computed from HIT distances as well. From these functions, $g_{\text{HIT}}(h)$, we derive correlation lengths $\ell_{\text{HIT}}$ in completely analogous manner. In general, HIT correlation lengths are shorter than tree correlation lengths, but, apart from some scaling function, they show essentially the same qualitative features as the tree correlation lengths and we dispense from details here which will be the subject of a forthcoming paper.

The whole procedure to compute tree distance autocorrelation functions can be carried over to free energies provided the tree distance is replaced by the absolute difference in free energies, $d_f(i,j)$ according to equation (12). We obtain thereby a new method to compute autocorrelation functions of free energies from density surfaces which represents an alternative to the random walk technique. The uniform sampling method seems to have the advantage that more distant classes of sequences are treated with higher numerical accuracy.
6. Comparison with natural sequences

In order to compare the statistical results computed for secondary structures of random RNA sequences with natural RNA sequences examples were chosen with a base distribution as close to the uniform distribution as possible \((n_A \approx n_U \approx n_G \approx n_C \approx \nu/4)\). A sample which meets this requirement consists of 12 full mature m-RNA molecules, i.e. with the introns removed, coding for \(\beta\)-globin molecules\(^{37}\) from different animals with chain lengths varying from 534 to 627. The sequence which deviates most strongly from the uniform distribution comes from \textit{xenopus laevis}:

\[
\frac{n_A}{n_U}/\frac{n_G}{n_C} = 0.29/0.26/0.21/0.24.
\]

The 12 sequences were folded and the five structures with lowest free energies were considered for statistical analysis (The sample thus contains 60 structures). The five structures span a energy band of about 1-2\% of the absolute free energy of the optimal structures. The mean number of base pairs per 100 nucleotides of the \(\beta\)-globin m-RNA sequences is 31. The corresponding quantity derived from random sequences of approximately the same chain length, \(\nu = 500\)\(^ {15}\), is 29.04 (For precise comparisons random sequences of approximately the same chain lengths have to be used since there is still some chain length dependence of the quantities in question. If we used, for example, random sequences of chain length \(\nu = 100\) the mean number of base pairs per 100 bases would be only 24.4).

In addition to \(\beta\)-globin m-RNAs, RNA molecules from other sources were considered as well: 14 eubacterial 16s r-RNAs\(^ {38}\) and 8 mitochondrial 16s r-RNAs\(^ {39}\). The ribosomal RNA molecules show substantially larger deviation from the uniform base distribution than the \(\beta\)-globin m-RNAs. For statistical analysis we choose again the minimum free energy structure together with the four most stable suboptimal folding patterns. The energy bands spanned by the five structures lies within a range of 1-2\% of the absolute free energy of the optimal structure, as it was found with the \(\beta\)-globin m-RNAs.
Table 2: Comparison of mean stack size, mean loop size and mean branching
degree of loops between secondary structures computed for random†
and natural RNA sequences‡.

<table>
<thead>
<tr>
<th>Source</th>
<th>Mean stack size</th>
<th>Mean loop size</th>
<th>Mean loop degree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{n}_{st}$</td>
<td>$\bar{n}_{lp}$</td>
<td>$\bar{n}_{LD}$</td>
</tr>
<tr>
<td>Random RNA sequences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUGC</td>
<td>4.57</td>
<td>5.42</td>
<td>1.82</td>
</tr>
<tr>
<td>AU</td>
<td>7.66</td>
<td>4.69</td>
<td>1.78</td>
</tr>
<tr>
<td>GC</td>
<td>6.46</td>
<td>2.98</td>
<td>1.92</td>
</tr>
<tr>
<td>Natural RNA sequences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta$-globin m-RNAs</td>
<td>4.49</td>
<td>4.42</td>
<td>1.89</td>
</tr>
<tr>
<td>mitochondrial r-RNAs</td>
<td>4.44</td>
<td>0.53</td>
<td>1.74</td>
</tr>
<tr>
<td>eubacterial r-RNAs</td>
<td>4.59</td>
<td>4.62</td>
<td>1.92</td>
</tr>
</tbody>
</table>

† About 50 000 structures from random sequences of chain length $v = 500$ with
different base pairing alphabets$^{15}$.

‡ Sample sizes: 12 $\beta$-globin m-RNAs$^{37}$, 14 mitochondrial 16s r-RNAs$^{38}$ and 8
eubacterial 16s r-RNAs$^{39}$. For each sequence the sample contains the min­
imum energy structure together with the four suboptimal foldings of lowest
free energies.

In table 2 mean stack sizes, mean loops sizes and mean branching degrees of
loops from the three natural samples are compared with those from random se­
quences. The data computed for AUGC random sequences with uniform nucleotide
distribution are complemented by those for pure AU- and pure GC-sequences in
order to provide information on the dependence of statistical properties on base
distributions. As expected the mean values obtained from $\beta$-globin m-RNAs fit
best the data from random sequences: firstly, these sequences are closest to the
uniform distribution of nucleotide bases, and secondly, m-RNAs are commonly
thought to have only few structural restrictions for proper function.

In order to make the comparison with the experimental data more precise we computed probability densities for stack sizes, loop sizes and loop branching degrees. The results are shown in figures 14, 15 and 16. In the case of stack size probability densities the agreement between the β-globin m-RNAs and the random sample of AUGC-sequences is very good. The data computed from the other two samples from r-RNAs fit the curve from the random sample not nearly as well.

**Fig. 14:** Probability densities \( P(n_{st}) \) of stack sizes in natural RNA sequences compared with random sequences of chain lengths \( \nu = 500 \). One curve shows the data computed for m-RNAs of β-globins (■), individual points are given for eubacterial 16S r-RNAs (●) and for mitochondrial 16S r-RNAs (∗). The second curve refers to large ensembles of random RNA sequences built from the AUGC alphabet with chain lengths \( \nu = 500 \) (○).

**Fig. 15:** Probability densities \( P(n_{lp}) \) of loop sizes in natural RNA sequences compared with random sequences of chain lengths \( \nu = 500 \). One curve shows the data computed for m-RNAs of β-globins (■), individual points are given for eubacterial 16S r-RNAs (●) and for mitochondrial 16S r-RNAs (∗). The second curve refers to large ensembles of random RNA sequences built from the AUGC alphabet with chain lengths \( \nu = 500 \) (○).

The probability densities of loop sizes for natural and random sequences are compared in figure 15. In essence, the results are the same as with the probability distribution for stack sizes: the data from m-RNAs of β-globins fit the curves computed for random AUGC-sequences of chain length \( \nu = 500 \) much better than the points obtained from the r-RNAs. In detail, however, the agreement between the β-globin m-RNAs and the random RNA sequences is not as good as for the stack sizes. The natural sequences have significantly more bulges of size 1 than
random sequences. In addition, loops of size 3 are more probable and those of size 4 less probable in random sequences than in the β-globin m-RNAs. This result might be consequence of a preference for especially stable tetraloops in the m-RNAs which were not considered in the random sample. Deviations at higher loop sizes presumably reflect scatter caused by the relatively small size of the natural sample.

**Fig. 16:** Probability densities $P(n_{LD})$ of degrees of branching in the loops of natural RNA sequences compared with those of random sequences of chain lengths $\nu = 500$. One curve shows the data computed for m-RNAs of β-globins (●), individual points are given for eubacterial 16S r-RNAs (●) and for mitochondrial 16S r-RNAs (★). The second curve refers to large ensembles of random RNA sequences built from the AUGC alphabet with chain lengths $\nu = 500^{15}$ (○).

Probability densities for the branching degree of loops (Fig.16) show again excellent agreement between β-globin m-RNAs and random sequences, and substantial deviations observed with the two samples derived from mitochondrial and eubacterial 16s r-RNAs.

The results of the comparison of secondary structures of natural and random sequences suggest an extension of the computations to RNA sequences with nucleotide distributions different from the -- most probable -- uniform distribution in order to be able to separate structural effects caused by the base distribution from those resulting from the function of the molecules. A forthcoming paper\textsuperscript{13} will be dealing in detail with this questions.
Acknowledgements

Financial support for the work reported here was provided by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (projects P 6864, P 8526 and S 5305), by the *Jubiläumsfonds der Österreichischen Nationalbank* (project no. 3819), by the Austrian *Bundesministerium für Wissenschaft und Forschung* (GZ 30.330/2-23/90), by the German *Volkswagen-Stiftung*, by the *John D. and Catherine T. MacArthur Foundation*, by the *National Science Foundation* (PHY-8714918) and by the U.S. *Department of Energy* (ER-FG05-88ER25054). An IBM 6000 RISC workstation was generously supplied by the *EDV-Zentrum der Universität Wien* within the EASI project of IBM. We are grateful to Dipl.Ings. Erich Bauer, Manfred Tacker and Ivo L. Hofacker for performing some computation and providing data. Useful hints in stimulating discussions given by Professors Doyne Farmer, Paulien Hogeweg, Stuart Kauffman, John McCaskill, and Alan Perelson are gratefully acknowledged. We thank Dr. Michael Ramek for providing *TeX* MACROS$^{40}$ for drawing diagrams.
References


37. Genbank names: gothbbaa, hsbgl1, hsdgl1, hsggl4, lebglob, mushbbmaj, ptggglog, rabhbb, rbbbb3, ratglbr, xebbeta and xlbglr.

38. Genbank names: anlmttggr, bovmt, ceumtfvla, frgmtrc12s, gotmttggr, hummtgc, hyrmtfvla, mmumtfvla, musmt, odmtfvla, palmtcg, ratmtrgpd, trgmttggr and xelmtrrza.

39. Genbank names: bacrgrrnb, deirgda, hclrgda, mpocpcg, m27040, prirrgda and stmrrnb.

\[ I_k = \text{AUGCGUUGGACGUGCAGCUCCAGUCAGCAGAUUGCUCAGUUGUUAAUUCGUGUGAGCGCGCUAGUCU} - \text{AGUCGGAAGGCGCGUCAGAUGUGCAAGCAUGUACGAAACGC} \]
Figure 3 of January 24, 1992
Figure 4 of January 24, 1992
Figure 5 of January 24, 1992
Figure 6 of January 24, 1992
Figure 7 of January 24, 1992
Figure 9 of January 24, 1992
Figure 10. GC (upper), AUGC (lower)
Figure 11. full trees (upper), HIT 0 (lower)
Figure 13 of January 24, 1992
Figure 14 of January 24, 1992
Figure 15 of January 24, 1992
Figure 16 of January 24, 1992