Statistics of RNA Secondary Structures [†]

By

Walter Fontana [b,c], Danielle A. M. Konings [d], Peter F. Stadler [a]

and Peter Schuster [a,c,e,*]

[a] Institut für Theoretische Chemie der Universität Wien

[b] Theoretical Division, Los Alamos National Laboratory

[c] Santa Fe Institute

[d] Department of Molecular, Cellular and Developmental Biology, University of Colorado at Boulder

[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) 40 480 678

Fax: (**43 1) 40 28 525

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

pks@imb-jena.de

Abstract.

A statistical reference for RNA secondary structures with minimum free energies is computed by folding large ensembles of random RNA sequences. Four nucleotide alphabets are used: two binary alphabets, **AU** and **GC**, the biophysical **AUGC** and the synthetic **GCXK** alphabet. RNA secondary structures are made of structural elements, such as stacks, loops, joints and free ends. Statistical properties of these elements are computed for small RNA molecules of chain lengths up to 100. The results of RNA structure statistics depend strongly on the particular alphabet chosen. The statistical reference is compared with the data derived from natural RNA molecules with similar base frequencies.
Secondary structures are represented as trees. Tree editing provides a quantitative measure for the distance, d_t, between two structures. We compute a structure density surface as the conditional probability of two structures having distance t given that their sequences have distance h. This surface indicates that the vast majority of possible minimum free energy secondary structures occur within a fairly small neighbourhood of any typical (random) sequence.
Correlation lengths for secondary structures in their tree representations are computed from probability densities. They are appropriate measures for the complexity of the sequence-structure-relation. The correlation length also provides a quantitative estimate for the mean sensitivity of structures to point-mutations.

Keywords.

Combinatory map – Correlation length – Landscape – Random RNA sequence – RNA secondary structure – Shape space covering – Tree editing

1. Introduction

A great variety of natural biopolymers were studied extensively by sequence analysis, X-ray diffraction and spectroscopic techniques. Their molecular structures are well known by now. In some cases the studies were extended to closely related variants differing in one or a few positions from the wild type sequences. Despite the availability of very detailed information on many individual biomolecules very little, if anything, is known about the statistics of structural features of biopolymers. General information on the sensitivity of structures against changes in the sequences is also very rare. In the case of RNA direct interest in the properties of molecules with random sequences is steadily growing: random RNA sequences are readily available by the current synthetic techniques and they are frequently used in applied molecular evolution [1–3]. A straightforward strategy for gaining knowledge about molecules with random sequences is to compute and analyze the structures of statistical ensembles of biopolymers. The computation of three-dimensional structures is, however, highly time consuming, and still encounters substantial theoretical and algorithmic difficulties.

The folding of RNA sequences into three-dimensional structures is decomposed into two steps:

(1) folding of the string of bases, into a secondary structure by base pair formation, and

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

The biophysical rationale for considering RNA secondary structures as a first crude approach to real RNA structures is based on several facts: RNA secondary structure formation covers the major part of the folding energy, RNA secondary structures can be used successfully to interpret RNA function and reactivity, and RNA secondary structures are conserved in evolution.

Secondary structures of RNA molecules, represented as the list of Watson-Crick (W-C)-like base pairs, are much easier to predict than three-dimensional structures. These structures are mainly determined by the conventional base pairing rules of RNA: (**G**≡**C**, **A**=**U** and **G**−**U**). In this paper we also consider the non-natural xanthine–2,6-diaminopyrimidine pair, **X**≡**K**, which was recently incorporated enzymatically into synthetic RNA and DNA molecules [4]. Base pairing and base pair stacking energies are generally larger than those of other interactions involved in the formation of spatial structures and arrange the RNA structure in helical segments interrupted by various types of loop regions. In the most narrow sense the term "secondary structure" is used to refer to the major list of W-C-like base pairs that constitute an unknotted structure which may be drawn as a planar graph. Thus, loop-loop interactions are not included in this definition. All the remaining W-C-like base pairs (e.g. those occuring in pseudoknots) and other types of interactions are then referred to as "tertiary interactions" within this concept. Although this classification is in a sense arbitrary, in particular with reference to pseudoknot structures [5, 6], it reflects the partitioning of the folding process into several stages, the first one being the nucleation of double-stranded helical regions. It is, therefore, meaningful to consider the folding into a secondary structure (even in this narrow sense) as a first step towards the formation of the full three-dimensional structure. The computational approach used here for the analysis of gross features of RNA secondary structures is based on this concept of "unknotted" secondary structures.

The predominantly used thermodynamic folding algorithm for RNA secondary structures was originally conceived by Zuker, Stiegler and Sankoff [7, 8], and is based on an application of dynamic programming to the RNA problem [9]. It was primarily designed to compute the minimum free energy structure, but derivative algorithms allow to obtain suboptimal foldings as well [10−12]. Alternatively one may consider suboptimal foldings with the corresponding Boltzmann weights and compute partition functions for RNA secondary structures directly [13]. The empirical parameters used in the folding algorithm were derived some years ago [14] (For a more recent updated version see ref. 15). Here we shall be dealing exclusively with minimum free energy secondary structures computed by a variant of the Zuker algorithm. The computer code was originally designed for fast folding as part of a simulation package for molecular evolution [16, 17]. The present version of the software package uses the updated version of the empirical parameter set [15], and includes a statistics program as well as tree editing routines. For the **XK** base pair we use the **GC** parameter set which seem to come closest to the base pairing strength of the synthetic base pair [4].

In this paper we are concerned with the statistical properties of RNA secondary structures and their dependence on various base alphabets: the natural four letter alphabet (**AUGC**), a non-natural **GCXK**-alphabet, and the two letter alphabets (**AU** and **GC**). The **GCXK**-alphabet provides an interesting case since it contains two base pairs of approximately equal strength, and it is free of complementarity violations (**G**−**U** in the natural alphabet). Comparison of structures derived from different alphabets allows to separate effects of different origin and helps to understand the complex superposition of contributions from the size of the alphabet, from the pairing rules and from the strength of base pairs in the natural RNAs.

The statistics of structural elements for small RNA molecules (with chain length $\nu < 100$) is presented in section 2. These data are considered as a statistical reference to be compared with data from natural sequences in section 3. In section 4 we present a quantitative measure for the distance between RNA secondary structures. Sequence-structure-relations are viewed as *combinatory maps* (CMs) from sequence space into a shape space. The notion of *structure density surfaces* (SDS) is introduced in section 5. The SDS casts statistical aspects of the relation between sequences and structures into a condensed form, and provides a tool to derive and calculate global properties for classes of RNA molecules. Autocorrelation functions and correlation lengths of structures finally characterize the sequence-structure-relation by a single function or a single number, respectively. They provide a useful measure for the sensitivity of RNA structures against point-mutations (section 6).

2. Statistics of elements of RNA secondary structures

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

$$\mathbf{I}_k = \{\mathsf{AUGCGUUGGACGUGCAGUCCAGUCAG} \ldots \mathsf{AAACGC}\} ,$$

into a secondary structure $S_k = \mathcal{S}(\mathbf{I}_k)$ where $\mathcal{S}(.)$ stands for the folding algorithm which computes a unique structure for every sequence $\mathbf{I}_k$. An example is shown in figure 1. Many widely different 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.

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 ($\circ$) and base pairs as a full circles ($\bullet$), respectively. A root ($\blacksquare$), 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.

A secondary structure is viewed conventionally as a combination of *structure elements* which fall into seven classes:

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).

Stacks stabilize secondary structures whereas loops have a destabilizing effect whose magnitude depends on the size of the loop. 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

$$n_{\mathrm{L}} = n_{\mathrm{H}} + n_{\mathrm{B}} + n_{\mathrm{I}} + n_{\mathrm{M}} \ , \tag{1}$$

which must be identical to the number of stacks, $n_{\mathrm{L}} = n_{\mathrm{S}}$. Nucleotides in joints and free ends are often termed *external* bases.

We generated a statistical reference for RNA secondary structures on sequences of length $20 \leq \nu \leq 100$ built from various alphabets of size κ ($\kappa = 2$: **AU**, **GC**, and $\kappa = 4$: **AUGC**, **GCXK**). Equal probability of digits was assumed and hence the base frequences of random sequences lie around the most probable distributions, (0.25,0.25,0.25,0.25) or (0.5,0.5), respectively. In this section we report on the average occurence of secondary structure elements.

Base pairs: The mean number of base pairs (n_{BP}) increases linearly with the chain length ν for sufficiently long sequences (Fig.2). Deviations at small chain lengths ($\nu < 50$) are found with **AU**-, **AUGC**-, and **GCXK**-sequences. The influence of the alphabet is readily interpreted by considering the *stickiness* P of the sequences which is understood as the probability that two arbitrarily chosen bases can form a base pair. For the uniform base compositions one has $P_{\mathsf{AU}} = P_{\mathsf{GC}} = 0.5$, $P_{\mathsf{AUGC}} = 0.375$, and $P_{\mathsf{GCXK}} = 0.25$. As expected, and as seen in figure 2, the pure **GC**-sequences are leading with respect in the number of base pairs since they have the highest possible stickiness and form the strongest base pairs. Destabilization by loops is more readily compensated by **GC** pairs than by **AU** pairs, and 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. The fact that **AUGC**- and **GCXK**-sequences have almost the same mean number of base pairs is fortuitous: the former are more sticky, the latter form stronger base pairs and the two effects cancel by accident.

Fig. 2: The mean number of base pairs (n_{BP}) as a function of the chain length ν. Values are shown for binary **GC**-sequences ($\diamond$), for binary **AU**-sequences ($\circ$), for four letter **GCXK**-sequences with **GC** parameters ($*$), and for natural **AUGC**-sequences ($\bullet$). Depending on the chain length ν the mean values are computed from samples of 50 000 ($\nu = 20$) to 500 000 sequences ($\nu = 100$) which were sampled by the technique described in section 5. Unstable structures, i.e. structures with non-negative free energies ($f \geq 0$) are not considered for structure statistics.

Fig. 3: The mean number of loops (n_{L}) or stacks (n_{S}) as a function of the chain length ν. Values are shown for binary **GC**-sequences ($\diamond$), for binary **AU**-sequences ($\circ$), for four letter **GCXK**-sequences with **GC** parameters ($*$), and for natural **AUGC**-sequences ($\bullet$). Computations were performed as described in figure 2.

Loops and stacks: The mean number of loops n_{L} per structure which is identical to the mean number of stacks n_{S} scales also linearly with chain length

ν (Fig.3). Weaker base pairing implies that closing of loops is more difficult and hence the structures derived from **AUCG**- or **AU**-sequences have fewer loops than their **GCXK** or **GC** counterparts. Sequences with lower stickiness values have on the average more loops than stickier sequences. The effect of base pair strength is apparently stronger than that of stickiness.

Fig. 4: The mean degree of loops ($n_{\rm LD}$) as a function of the chain length ν. Values are shown for binary **GC**-sequences ($\diamond$), for binary **AU**-sequences ($\circ$), for four letter **GCXK**-sequences with **GC** parameters ($*$), and for natural **AUGC**-sequences ($\bullet$). Computations were performed as described in figure 2.

Loop degree: The average degree of loops (Fig.4) is in the range $1 < n_{\rm LD} < 2$. It converges to a constant value with increasing chain length ν. Structures derived from sequences with strong base pairs (**GC**, **GCXK**) have more higher order branches than those obtained from **AUGC**- and **AU**-sequences.

Fig. 5: The mean number of base pairs in one stack ($n_{\rm st}$) as a function of the chain length ν. Values are shown for binary **GC**-sequences ($\diamond$), for binary **AU**-sequences ($\circ$), for four letter **GCXK**-sequences with **GC** parameters ($*$), and for natural **AUGC**-sequences ($\bullet$). Computations were performed as described in figure 2.

Fig. 6: The mean number of bases in one loop ($n_{\rm lp}$) as a function of the chain length ν. Values are shown for binary **GC**-sequences ($\diamond$), for binary **AU**-sequences ($\circ$), for four letter **GCXK**-seqeunces with **GC** parameters ($*$), and for natural **AUGC**-sequences ($\bullet$). Computations were performed as described in figure 2.

Stack sizes and loop sizes: Mean stack sizes ($n_{\rm st}$) and mean loop sizes ($n_{\rm lp}$) converge to almost constant values at fairly small chain lengths ($\nu \approx 50$) as shown in figures 5 and 6. Apparently the convergence of $n_{\rm st}$ and $n_{\rm lp}$ in structures of **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, on average longer in the **AU**-case than in the **GC**-case. Large loops are favoured by both low stickiness and weak base pairs since a weak closing pair is more likely to be destabilzed by a loop than a strong one. Thus, on average, **GC**-sequences form the smallest loops.

Components: Components are substructures which are connected by joints. The mean number of components $n_{\rm C}$ shows a characteristic lag phase before it starts off to increase with increasing chain length ν. This lag phase reflects the fact that a certain minimum chain length is required in order to form local substructures. The lag phase is more pronounced in structures built from alphabets with weaker base pairs (**AUGC**, **AU**). The increase of $n_{\rm C}$ with ν is much stronger in the case of the four letter alphabets. The data shown in figure 7 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 [18]. These computations have shown, however, that the chain length dependence of the number of components is not yet linear at chain lengths around $\nu \approx 500$. **GC**-sequences are special in a way since there the mean number of components converges to a value in the range $2.75 \le n_{\rm C} \le 3$.

Fig. 7: The mean number of components ($n_{\rm C}$) connected by $n_{\rm J} = n_{\rm C} - 1$ joints as a function of the chain length ν. Values are shown for binary **GC**-sequences ($\diamond$), for binary **AU**-sequences ($\circ$), for four letter **GCXK**-sequences with **GC** parameters ($*$), and for natural **AUGC**-sequences ($\bullet$). Computations were performed as described in figure 2.

Fig. 8: The mean number of unpaired external bases (n_{ext}) as a function of the chain length ν. Values are shown for binary **GC**-sequences ($\diamond$), for binary **AU**-sequences ($\circ$), for four letter **GCXK**-sequences with **GC** parameters ($*$), and for natural **AUGC**-sequences ($\bullet$). Computations were performed as described in figure 2.

External digits: The mean number of external bases (n_{ext}) shows a rather complex increases with the chain length ν. It contains two contributions, the number of terminal free bases and the number of free bases in joints. The number of joints as we saw before shows a non-linear increase with the chain length in the range in question (Fig.7). The strength of base pairs has a strong influence on n_{ext}: weak pairing implies larger mean numbers of external digits. Interestingly n_{ext} is almost constant for **GC**-sequences where the number of components converges to a constant value for longer sequences ($\nu > 300$) too [18].

So far we focussed only on mean values. Resolved distributions for stack sizes, loop sizes and loop degrees will be presented in the next section, where data from the statistical reference will be compared with those obtained from natural sources with similar base compositions.

3. Comparison with natural sequences

In order to compare data from the statistical reference for RNA secondary structures with those from natural RNA sequences, samples with base distributions as close as possible to the uniform distribution, $(n_{\mathsf{A}}, n_{\mathsf{U}}, n_{\mathsf{G}}, n_{\mathsf{C}}) \approx \nu(0.25, 0.25, 0.25, 0.25)$, are required. In addition we expect best agreement when the biological function of the RNA demands as little secondary structure as possible. A sample which meets both requirements consists of 12 full mature M-RNA molecules, i.e. with the introns removed, coding for β-globin molecules [19] from different animals. Within this sample the chain lengths vary from 534 to 627, the sequence which deviates most strongly from the uniform distribution comes from *xenopus laevis*: $(n_{\mathsf{A}}, n_{\mathsf{U}}, n_{\mathsf{G}}, n_{\mathsf{C}}) \approx \nu(0.29, 0.26, 0.21, 0.24)$. 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. Some mean values are shown in table 1 together with the statistical data for the **AUGC**-, the **AU**- and the **GC**-alphabet. The mean number of base pairs per 100 nucleotides of the β-globin M-RNA sequences is 31. It is slightly larger than the corresponding quantity derived from the statistical reference at approximately the same chain length, $\nu = 500$ [18]: 29.0 base pairs. The β-globin M-RNAs are slightly more stable than the statistical reference in the sense that they have lower free energies in their minimum free energy structures. The stability argument has been used previously [20] in sequence and structure comparisons of viriods and virusoids which, as commonly assumed, have been selected for stability by evolution. There the thermodynamically-calculated structures from natural sequences form indeed more base pairs than the corresponding sample of random sequences, for example 35 base pairs per 100 nucleotides in the CSV viroid RNA as compared to 29.5 in the random standard. The mean loop size (Table 1) is somewhat smaller in the β-globin M-RNA sample than in the reference. The mean branching degrees of loops and the mean stack sizes of the two samples agree well.

In addition to β-globin M-RNAs, two samples of natural RNA molecules from other sources were considered as well: 14 eubacterial 16s R-RNAs [21] and 8 mitochondrial 16s R-RNAs [22]. These two samples fulfil the two conditions for a meaningful comparison with the statistical reference less well: the ribosomal RNAs show substantially larger deviation from the uniform base distribution than the β-globin M-RNAs and their secondary structures are certainly important for biological function. For statistical analysis we choose again the minimum free

Table 1: Comparison of mean number of base pairs, mean stack size, mean loop size and mean branching degree of loops between secondary structures from the statistical reference † and from samples of natural RNA sequences ‡.

Source	$\overline{n}_{\mathrm{BP}}$ / ν	$\overline{n}_{\mathrm{st}}$	$\overline{n}_{\mathrm{lp}}$	$\overline{n}_{\mathrm{LD}}$
Statistical reference of RNA sequences				
AUGC	0.290	4.57	5.42	1.82
AU	0.354	7.66	4.69	1.78
GC	0.403	6.46	2.98	1.92
Natural RNA sequences				
β-globin M-RNAs	0.31	4.49	4.42	1.89
mitochondrial R-RNAs	0.25	4.44	6.53	1.74
eubacterial R-RNAs	0.32	4.59	4.62	1.92

† A sample of about 50 000 structures from sequences of chain length $\nu=500$ with different base pairing alphabets computed according section 5 [18].

‡ Sample sizes: 12 β-globin M-RNAs [19], 14 mitochondrial 16s R-RNAs [21] and 8 eubacterial 16s R-RNAs [22]. For each sequence the sample contains the minimum energy structure together with the four suboptimal foldings of lowest free energies.

energy structure together with the four most stable suboptimal folding patterns (which were again in the 1-2% energy range, as with the β-globin M-RNAs). As far as the mean values are concerned the eubacterial R-RNA sample fits the reference data as well as the M-RNA sample. There are however significant differences with the mitochondrial R-RNAs: the preference for larger loops is remarkable, and it is reflected by the smaller number of base pairs.

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 9 - 11. For stack size frequencies the agreement between the β-globin M-RNAs and the statistical reference from **AUGC**-sequences is very good. The data computed from the other two samples from R-RNAs fit the curve for the reference not nearly as well.

Fig. 9: Probability densities $P(n_{\mathrm{st}})$ of stack sizes in natural RNA sequences compared with the statistical reference of chain lengths $\nu = 500$. One curve shows the data computed from minimum free energy structures for M-RNAs of β-globins ($\bullet$), individual points are given for eubacterial 16S R-RNAs ($\times$) and for mitochondrial 16s R-RNAs ($+$). Further points are given for the phylogenetically-derived secondary structures [23] of the eubacterial ($\otimes$) and the mitochondrial ($\oplus$) R-RNAs. The second curve refers to the statistical reference built from the **AUGC** alphabet with chain lengths $\nu = 500$ [18] ($\circ$).

Fig. 10: Probability densities $P(n_{\mathrm{lp}})$ of loop sizes in natural RNA sequences compared with the statistical reference of chain lengths $\nu = 500$. One curve shows the data computed from minimum free energy structures for M-RNAs of β-globins ($\bullet$), individual points are given for eubacterial 16s R-RNAs ($\times$) and for mitochondrial 16S R-RNAs ($+$). The second curve refers to the statistical reference built from the **AUGC** alphabet with chain lengths $\nu = 500$ [18] ($\circ$).

The distribution of loop sizes for natural sequences are compared with the statistical reference (**AUGC**, $\nu = 500$) in figure 10. 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 the statistical reference 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 data for natural sequences show much larger deviations. Natural sequences have significantly more loops of size 1 and less loops of size 3 than the reference. The samples of natural sequences are inevitably small and hence the deviations at higher loop sizes need not be statistically significant. Apparently, loop size is a more sensitive parameter than stack size.

Fig. 11: Probability densities $P(n_{\mathrm{LD}})$ of degrees of branching in the loops of natural RNA sequences compared with the statistcal reference of chain lengths $\nu = 500$. One curve shows the data computed from minimum free energy structures for M-RNAs of β-globins ($\bullet$), individual points are given for eubacterial 16s R-RNAs ($\times$) and for mitochondrial 16s R-RNAs ($+$). The second curve refers to the statistical reference built from the **AUGC** RNA sequences built from the **AUGC** alphabet with chain lengths $\nu = 500$ [18] ($\circ$).

Frequency counts for the branching degree of loops (Fig.11) show excellent agreement between β-globin M-RNAs and random sequences. Substantial deviations are observed only with the two samples derived from mitochondrial and eubacterial 16s R-RNAs.

Phylogenetically-derived structures of RNA molecules are considered biologically more significant than minimum free energy structures. For the purpose of comparison we analyzed also a sample of phylogenetically-derived 16s R-RNA structures [23] and included the data for the stack size distribution in figure 9. We observe much larger scatter from the statistical reference than for the data points from minimum free energy structures. One feature seems to be of special interest: the phylogenetic 16s R-RNA structures show a significant local peek for a stack size of seven base pairs which does neither exist in the minimum free energy samples nor in the phylogenetic data for 23s R-RNAs (not shown here). This is a clear hint that there is an unusual favour of stacks with seven base pairs in the 16s R-RNA s, and accordingly these stacks may have a specific function. Loop size and loop degree distributions of phylogenetic structures show also larger scatter than the minimum free enery samples of the natural sequences.

The results obtained in the comparison of secondary structures of natural sequences with the statistical reference suggest to propose this procedure as a tool for testing biological function. If a sufficiently large sample of natural sequences shows detailed agreement with the statistical reference of the same base composition then it is very unlikely that biological function imposes a severe constraint on structures. Since the same folding algorithm is used in the computation of the reference and the sample to be compared, the empirical parameters and the particular folding routine chosen have only minor influence on the result. The comparison is done on the level of ensembles and hence individual errors are fairly unimportant.

4. Distances between structures

RNA secondary structures can be represented as trees [8, 24, 25]. A secondary structure S_k is converted one to one into a tree T_k by assigning an internal node to each base pair and a leaf node to each unpaired digit (Fig.1). The conversion starts with a root which does not correspond to a physical unit of the RNA molecule. It is introduced to prevent the formation of a tree forest for RNA structures with external elements. For details of the interconversion of secondary structures and trees see table 2.

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 in the HIT representation.

Table 2: Interconversion of secondary structures and trees

Accessibility of bases and base pairs [8]:
The bases are labeled with their position, i, in the sequence. Consider a base pair $i-j$ with $i<j$ in the secondary structure graph. A base r is said to be *accessible* from the base pair $i-j$, if there is no base pair $x-y$ such that $i<x<r<y<j$. A base pair $r-s$ is said to be accessible if r and s are accessible from the base pair $i-j$.

A secondary structure graph (SSG) is converted into a tree graph (TG) by the following procedure:
(1) Assign to each unpaired base a leaf node, and to each base pair (two SSG-nodes) one internal TG-node.
(2) The parent of each TG-node is the internal node corresponding to that accessible base pair $k-l$ $(k<l)$ which contains the smallest k. The parent of nodes corresponding to external bases and external base pairs is a virtual node representing the root of the tree.
(3) Connect each node to its parent.

A TG is converted into a SSG by the following procedure:
(1) Replace each internal TG-node by a connected pair of SSG-nodes (base pair). The TG-edges are inherited by the left SSG-node of the pair.
(2) For any TG-node with a left sibling replace the edge to its parent with an edge to its left sibling.
(3) An SSG-node that is base paired has three edges (two from the backbone, and one from the pairing), otherwise it has two. 5'- and 3'-ends have one less. Complete the SSG by inserting all missing edges at a node as connections to the corresponding parent in the TG, proceeding from deep to shallow levels.

The transformation from the full tree to the HIT retains complete information on the structure. Secondary structure, full tree as well as HIT are equivalent:

$$\mathbf{I}_k \Longrightarrow S_k \Longleftrightarrow T_k \Longleftrightarrow H_k \ . \tag{2}$$

Tree editing induces a distance in the space of trees, and hence also in the space of RNA secondary structures. A tree is transformed into another tree by a series of editing operations with predefined costs [24–27]. The distance between two trees,

$$d(T_j, T_k) \; = \; d_t(j,k) \; , \tag{3}$$

is the smallest sum of the costs along an editing path. It can be shown that this tree distance forms a metric in the space of trees [27]. The parameters used in our tree editing are summarized in table 3. The editing operations preserve the relative traversal order of the tree nodes. Tree 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 3 for editing operations on weighted trees, distances between HITs can be computed as well:

$$d(H_j, H_k) \; = \; d_t^{(HIT)}(j,k) \; . \tag{4}$$

Tree distances between full trees and HITs fulfil the general relation:

$$d_t \; \leq \; d_t^{(HIT)} \; . \tag{5}$$

A proof of the inequality (5) will be given elsewhere.

An alternative graphical method for the comparison of RNA secondary structures [28–30] encodes secondary structures as linear strings with balanced parentheses representing the base pairs, and some other symbol coding for unpaired positions. Distances are computed by direct end-to-end alignment of the strings.

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 previously to solve this problem [24, 25]. They are based on plausible *ad hoc* assumptions. The

Table 3: 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.

Edit operation	Symbolic notation	Edit parameters Symbol	Cost
Full Tree Edit Operations			
Single base insertion	$0 \longrightarrow \circ$	$\delta_\circ$	1
Base pair insertion	$0 \longrightarrow \bullet$	$\delta_\bullet$	2
Single base deletion	$\circ \longrightarrow 0$	$\delta_\circ$	1
Base pair deletion	$\bullet \longrightarrow 0$	$\delta_\bullet$	2
Relabel		ρ	0
Substitution	$\bullet \longleftrightarrow \circ$	σ	1
HIT Edit Operations			
Insertion of single bases	$0 \longrightarrow \circ\, w$	$w \cdot \delta_\circ$	w
Insertion of base pairs	$0 \longrightarrow \bullet\, w$	$w \cdot \delta_\bullet$	$2w$
Deletion of single bases	$w \circ \longrightarrow 0$	$w \cdot \delta_\circ$	w
Deletion of base pairs	$w \bullet \longrightarrow 0$	$w \cdot \delta_\bullet$	$2w$
Relabel of single bases	$w \circ \longleftrightarrow \circ\, v$	$\|w-v\| \cdot \delta_\circ$	$\|w-v\|$
Relabel of base pairs	$w \bullet \longleftrightarrow \bullet\, v$	$\|w-v\| \cdot \delta_\bullet$	$2\,\|w-v\|$
Substitution	$w \bullet \longleftrightarrow \circ\, v$		
	$\sigma \ \min\{w,v\} + \|w-v\| \cdot \begin{cases} \delta_\bullet & \text{if } w > v \\ \delta_\circ & \text{if } w < v \end{cases}$		

HITs introduced here are particularly well suited for unambiguous coarse-graining. The weights w allow straightforward definition of a resolution parameter: all nodes up to some predefined value, $0 < w \leq w_r$, are omitted, and the secondary structure corresponding to the coarse grained tree thus contains only structural features above a certain size.

5. Landscapes, combinatory maps, and density surfaces

The notion of *landscapes* was introduced into biology by the pioneering paper of Sewall Wright [31]. He visualized evolution as an optimization process on a *fitness landscape.* Since fitness is an exceedingly complex quantity, and since the occurence of unconstrained optimization is at least debatable, Wright's metaphor turned out to be not more than a heuristic principle. The landscape concept in biology saw a recent revival in model studies of evolution and coevolution [32–34], in theoretical immunology [35, 36], and in studies based on RNA structures and properties [16, 17, 37].

Comparison of RNA sequences is standard in molecular evolution. The distance between two sequences of the same length ν is given by the Hamming metric $d_h(k,\ell)$ [38, 39] which counts the number of different digits in the two end-to-end aligned sequences $\mathbf{I}_k$ and $\mathbf{I}_\ell$. The set of all sequences of a given length ν is of combinatorial complexity: an alphabet with κ letters yields κ^ν 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* [38, 40] for which d_h is a metric [39]. The sequence space of binary sequences ($\kappa = 2$) is a hypercube of dimension ν. In the case of four letter alphabets ($\kappa = 4$) the sequence space is more complex and can be understood as a union of two hypercubes with the appropriate connections. An earlier variant of sequence space is the *protein space* introduced by John Maynard-Smith [41].

A landscape is obtained by assigning a scalar quantity, for example a fitness value, a free energy or a rate constant, to every point in sequence space. Thus it is a mapping of the sequence space into the real numbers:

$$\mathcal{L}: \quad (X, d_h) \quad \Longrightarrow \quad (\mathbb{R}^1) \; . \tag{6}$$

In order to point at the combinatorial complexity of the support we refer to such an object as a *combinatory landscape* (CL).

Secondary structures – or trees – are as well elements from a space of combinatorial complexity with the tree distance $d_t(k,\ell)$ as a metric. The notion of a *shape space*, originally conceived for antibody-antigen recognition in theoretical immunology[42, 43], appears to be a useful concept for RNA secondary structures as well. The process of folding assigns to an element $\mathbf{I}_k$ of the sequence space a tree representing a secondary structure S_k. RNA folding, therefore, is a mapping Φ from a sequence space X into a shape space Y:

$$\Phi : \quad \big(X, d_h\big) \quad \Longrightarrow \quad \big(Y, d_t\big) \ . \tag{7}$$

The elements of both spaces are discrete structures of combinatorial complexity. Both spaces are endowed with appropriate metrics. We refer to a mapping of this kind as a *combinatory map* (CM).

As a generic tool for the study of CMs we propose the use of a two dimensional probability density $P(.,h)$. A tree distance density surface, for example, expresses the joint probability of two sequences of length ν, $\mathbf{I}_i$ and $\mathbf{I}_j$, to have Hamming distance $d_h(i,j)=h$, as well as folding into secondary structures with tree distance $d_t(i,j)=t$. The surface $P(t,h)$ is biassed along the Hamming distance axis with the distribution

$$p(h) \ = \ \binom{\nu}{h} (\kappa-1)^{\,h}\, \kappa^{\,-\nu} \,, \tag{8}$$

since there are $\kappa^{\nu}\cdot p(h)$ sequences at Hamming distance h from a given sequence. 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:

$$P(t|h) \ = \ \frac{P(t,h)}{p(h)} \ \approx \ \frac{n(t,h)}{\sum_{t=0}^{t_{\max}} n(t,h)} \ . \tag{9}$$

As indicated in equation (9) 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; $t_{\max}$ is used for the maximal tree distance. The total sample size is $N = \sum_{t=0}^{t_{\max}} \sum_{h=0}^{\nu} n(t,h)$, and from the definition of conditional probalities it follows that $\sum_{t=0}^{t_{\max}} P(t|h) = 1$ for each h.

We use a sampling technique that computes directly the 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 ℓ sequences at each Hamming distance $h = 1, 2, \ldots, \nu$ from the reference sequence,

(iii) we fold these sequences into secondary structures and compute tree distances relative to the structure of the reference sequence,

(iv) we bin them in (t,h) boxes by counting numbers of instances, and

(v) continue with item (i) until convergence, or some desired accuracy, has been achieved after, say, r reference points.

The sampling procedure corrects directly for the bias of the binomial distribution. By definition we have

$$\sum_{t=0}^{t_{max}} n(t,h) \;=\; r \cdot \ell \;=\; \text{const.} \;, \tag{10}$$

The sample size now is $N = r \cdot \ell \cdot \nu$. For binary sequences a slight modification of this procedure is required since we have only a single sequence at Hamming distance ν.

Fig. 12: Probability density surfaces for tree distances of secondary structures of **AUGC**- (**A**) and **GC**-sequences (**B**) of chain length $\nu = 100$. In this computation $r = 1000$ reference points and a sample of $\ell = 10$ sequences were used which amounts to a total sample size of 10^6.

The tree distance probability surfaces represent the average distribution of tree distances from a reference sequence. In the examples shown in figure 12 only 1000 reference sequences and 10^6 data points were already sufficient to yield a roughly constant surface. Convergence of this sampling technique is remarkably fast. Both surfaces from figure 12 show an overall shape and, superimposed upon it, rich and bizarre looking details. The overall shape of density surfaces for binary sequences (e.g. the **GC**-surface in fig.12) is horseshoe-like, and exhibits mirror-plane-symmetry. This symmetry is only approximate. In the binary case a sequence at distance ν from the reference has a nucleotide arrangement that is clearly compatible with the secondary structure of the reference. As a matter of fact such a sequence indeed folds into the reference structure or into a quite related one. Since for binary alphabets there is only one sequence at Hamming distance ν from a reference, this symmetry shows up. In contrast, **AUGC**-sequences at distance ν from a reference sequence need no longer be compatible with the reference structure. The overall shape of the tree density surface of **AUGC**-sequences, therefore, looks like one half of a horseshoe. The first half of the density surface ($0 \leq h \leq \frac{\nu}{2}$) resembles closely the corresponding part of the surface of binary sequences. The second half ($\frac{\nu}{2} \leq h \leq \nu$), however, is like a mountain ridge parallel to the Hamming axis, and indicates that the distribution of tree distances is more or less independent of the Hamming distance h.

Fig. 13: 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 does not depend on the usage of full trees. Essentially the same features are observed with density surfaces whose structure distances are computed from HITs (Fig.13). Moreover, the shape of the density surface is not altered when the structure distance measure is based on the above mentioned linear encoding of secondary structures [44]. The overall shape of distance density surfaces is a robust statistical feature of RNA secondary structure folding. It is, therefore, appropriate to use the generic term *structure density surface* (SDS) for this object.

First we interpret the shape of the first half of the SDS $(0 \leq h \leq \frac{\nu}{2})$ which is common to two-letter and four-letter alphabets. At very small Hamming distances from the reference sequence $(d_h = h = 1, 2, 3)$ the most probable structures are similar to that of the reference, and the probability density $P(t|h)$ has its maximum at small tree distances. With increasing Hamming distance this maximum shifts towards larger tree distances, and the probability to find the same or a closely related structure is already very small at intermediate Hamming distances $(4 \leq h \leq h_{cr})$. For Hamming distances $h \geq h_{cr}$ the probability densities are very similar and, apart from finer details, independent of h. The critical Hamming distance lies at the (more or less sharp) turn of the horseshoe and represents the change from local to global features of the SDS. Thus, the value of h_{cr} defines a neighborhood size around an average sequence within which secondary structures can be effectively randomized. An extension of the interpretation of SDS to the second half of the sequence space $(\frac{\nu}{2} \leq h \leq \nu)$ is straightforward by the arguments on structure compatible sequences given in the last but one paragraph.

Shape space covering: Since the above mentioned aspect of density surfaces does neither depend on the chain length ν, 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 as minimum free energy structures on random sequences, occur already within a small neighborhood of any random sequence. This conjecture has been strengthened by further independent techniques on which we will report elsewhere.

Probability density surfaces can be computed for any quantity for which a measure of distance exists. For example, a free energy density surface may be constructed and computed by substituting the absolute value of the difference in free energies,

$$d_f(i,j) = |f(S_i) - f(S_j)| = f \ , \tag{11}$$

for the tree distance $d_t(i,j)$. All other scalar properties of RNA molecules can be studied analoguously.

6. Autocorrelation functions and correlation lengths

Landscapes and CMs in general can be characterized statistically by autocorrelation functions [40, 45], which can be expressed in terms of mean square distances:

$$\varrho(h) \ = \ 1 - \frac{< d^2(h) >}{< d^2 >} \ . \tag{12}$$

$< d^2 >$ is the mean square distance sampled over the entire sequence space, and the mean square distance conditioned on sequences with Hamming distance h is denoted by

$$< d^2(h) > := < d^2\big(i,k \mid d_h(i,k) = h\big) > \ . \tag{13}$$

The autocorrelation function of tree distances, for example, can be computed from the density surface $P(t|h)$. The conditional mean square distance is simply the expectation value of t^2 computed for a given Hamming distance h

$$< d_t^2(h) > = \sum_{t=0}^{t_{\max}} t^2 \, P(t|h) \ \approx \ \frac{\sum_{t=0}^{t_{\max}} t^2 \, n(t,h)}{\sum_{t=0}^{t_{\max}} n(t,h)} \ . \tag{14}$$

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

$$< d_t^2 > = \sum_{h=0}^{\nu} < d_t^2(h) > \cdot \, p(h) \ \approx \ \sum_{h=0}^{\nu} \frac{\sum_{t=0}^{t_{\max}} t^2 \, n(t,h) \, p(h)}{\sum_{t=0}^{t_{\max}} n(t,h)} \ . \tag{15}$$

For uniform sampling, $\sum_{t=0}^{t_{max}} n(t,h) = \ell.r =$ const., the autocorrelation function can now be written in terms of elements of the sampling array $n(t,h)$,

$$\varrho_t(h) \;=\; 1 - \frac{\sum_{t=0}^{t_{\max}} t^2\, n(t,h)}{\sum_{h=0}^{\nu} \sum_{t=0}^{t_{\max}} t^2\, n(t,h)\, p(h)} \;. \tag{16}$$

Equation (16) is directly applicable during the computation of the density surface and hence convergence of the autocorrelation function can be monitored at runtime.

Computation of autocorrelation functions from density surfaces by means of equation (12) represents an alternative to the random walk technique [37, 45]. Since the more distant mutant classes are treated with higher numerical accuracy by uniform sampling this method has an advantage over random walks. In practice convergence of the computed autocorrelation functions is faster in the uniform sampling case.

Fig. 14: Correlation lengths of tree distances (ℓ_t) of RNA molecules in their most stable secondary structures as functions of the chain length ν. Values are shown for binary **GC**-sequences ($\diamond$), for binary **AU**-sequences ($\circ$), for four letter **GCXK**-sequences with **GC** parameters ($*$), and for natural **AUGC**-sequences ($\bullet$). Correlation lengths are calculated from $(\ln\, \varrho_t(h), h)$-plots by means of a least root mean square deviation fit.

Autocorrelation functions of tree distances $\varrho_t(h)$ are approximated by an exponential fit in order to calculate a *correlation length* (ℓ_t) for secondary structures in sequence space: $\ln\, \varrho_t(\ell_t) \;=\; -1$. As we conclude from figure 14 the correlation length increases roughly linearly with the chain length ν. It depends strongly on the base pairing alphabet. The correlation lengths of structures from binary sequences, **AU** or **GC**, are very similar and substantially shorter than those derived from **GCXK**-sequences and **AUGC**-sequences. There seems to be a trend towards longer correlation lengths with weaker base pairing: for sequences of chain lengths

$\nu \geq 30$ the correlation lengths for **AU** are slightly longer than those for **GC**, and the ℓ_t-values for **AUGC** are always longer than the values for **GCXK**.

The correlation length represent a global charcterization of a CM by a single value. The shorter the correlation length, the more complicated the sequence-structure-relation, and optimization of properties related to structure occurs on a very rugged landscape. The correlation length provides a measure of the mean stability of structures against point-mutations in their sequences. If the correlation length is long a sequence will tolerate on average many point-mutations without changing its minimum free energy structure. Apparently, natural sequences are least sensitive to point-mutations. This sensitivity can be changed by variation of the relative **GC**-content of sequences: the structures of sequences with predominantly one base pair have lower stability against point-mutations. An observation of extensive structural variations in 28s R-RNAs from different vertebrates may be related to this lack in resilience against point-mutation: **GC**-rich phylogenetically variable regions show large scale structural variations as detected by electron microscopy [46].

7. Concluding remarks

A novel approach to investigate biopolymer structures which does not focus on the properties derived from single sequences has been proposed and carried out for RNA secondary structures. A statistical reference is computed for structures with given chain lengths and base paring alphabets. It provides infomation on the statistics of structural elements. This statistics is complemented by structure density surfaces (SDS) which describe local and global features of sequence-structure-relations in condensed form. Correlation lengths of secondary structures in sequence space provide a measure for the predictability of structural changes as sequences are modified by point-mutations.

Although a particular empirical parameter set had to be used in the computation, most of the features discussed here are robust with respect to corrections in individual parameters. What does depend on the choice of parameters are, of course, the predictions of structures for single sequences as well as some numerical values of the statistical reference. For example, the frequencies of structural elements are sensitive to the parameters determining the stability of these elements. The results of the comparison with natural sequences (whose structures were computed with the same folding routine as the statistical reference), the shapes of the structure density distributions, or the correlation lenghts and their dependence on chain lengths and base pairing alphabets will be largely independent of parameters and other details of the folding algorithms. This has actually been tested in one particular case: consideration of especially stable tetraloops did not significantly change correlation lengths.

Our approach yields three major results:

(1) Comparison of sequence data at the level of structures computed for sufficiently large samples with the statistical reference is proposed as a tool for the detection of structural features determined by biological function.

(2) Most typical structures are found in close neighborhoods of any random sequence. To find sequences that fold into predetermined structures only regions of a small size in sequence space have to be searched, and any random sequence is an equally valid starting point. The size of this region can be read off directly from the SDS.

(3) A comparison of correlation lengths for secondary structures in sequence space shows that structures derived from **GC**- or **AU**-rich sequences are much more sensitive against mutation than those from **AUGC**-sequences with uniform base distributions.

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. Mac Arthur Foundation*, by the *National Science Foundation* (PHY-8714918) and by the U.S. *Department of Energy* (ER-FG05-88ER25054). D.A.M.Konings thanks the W.M.Keck Foundation for its generous support of RNA science on the Boulder campus. 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 Dr.R.R. Gutell for providing the ribosomal RNA structure files, and to Dipl.Ings. Erich Bauer, Manfred Tacker and Ivo L. Hofacker for performing some computations referred to here. Useful hints in stimulating discussions given by Professors Doyne Farmer, Paulien Hogeweg, Stuart Kauffman, Ted Maden, John McCaskill, Alan Perelson, and Detlef Riesner are gratefully acknowledged. Dr.Michael Ramek provided TEX MACROS [47] for drawing diagrams.

References

1. Horwitz, M.S.Z, Dube, D.K. & Loeb, L.A. (1989) *Genome* **31**, 112-117.

2. Tuerk, C. & Gold, L. (1990) *Science* **249**, 505-510.

3. Ellington, A.D. & Szostak, J.W. (1990) *Nature* **346**, 818-822.

4. Piccirilli, J.A., Krauch, T., Moroney, S.E. & Benner, S.A. (1990) *Nature* **343**, 33-37.

5. Pleij, C.W.A. (1990) *Trends Biochem.Sci.* **15**, 143-147.

6. Westhof, E. & Jaeger, L. (1992) *Current Opinion in Stuct.Biol.* **2**, 327-333.

7. Zuker, M. & Stiegler, P. (1981) *Nucl.Acids Res.* **9**, 133-148.

8. Zuker, M. & Sankoff, D. (1984) *Bull.Math.Biol.* **46**, 591-621.

9. Waterman, M.S. & Smith, T.F. (1978) *Math.Biosc.* **42**, 257-266.

10. Williams, A.L. & Tinoco, I. (1986) *Nucl.Acids Res.* **14**, 299-315.

11. Zuker, M. (1989) *Science* **244**, 48-52.

12. Jaeger, J.A. Turner, D.H. & Zuker, M. (1990) *Methods in Enzymology* **183**, 281-306.

13. McCaskill, J.S. (1990) *Biopolymers* **29**, 1105-1119.

14. 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**, 9373-9377.

15. Jaeger, J.A., Turner, D.H. & Zuker, M. (1989) *Proc.Natl.Acad.Sci.USA* **86**, 7706-7710.

16. Fontana, W. & Schuster, P. (1987) *Biophys.Chem.* **26**, 123-147.

17. Fontana, W., Schnabl, W. & Schuster, P. (1989) *Phys.Rev.A* **40**, 3301-3321.

18. Hofacker, I.L., Fontana, W., Konings, D.A.M. & Schuster, P. (1992) *Comparison of RNA secondary structures from natural and random samples.* In preparation.

19. Genbank names: gothbbaa, hsbgl1, hsdgl1, hsggl4, lebglob, mushbbmaj, ptggglog, rabhbba, rabhbb3, ratglbr, xebbeta and xlbgl1r.

20. Steger, G., Hofmann, H., Förtsch, J., Gross, H.J., Randles, J.W., Sänger, H.L. & Riesner, D. (1984) *J.Biomolec.Struct.Dynamics* **2**, 543-571.

21. Genbank names: anlmttgrg, bovmt, ceumtfvla, frgmtrc12s, gotmttgrg, hummtgc, hyrmtfvla, mmumtfvla, musmt, odomtfvla, palmtcg, ratmtrgpd, trgmttgrg and xelmtrrza.

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

23. Gutell, R.R. (1992) *Evolutionary characteristics of 16s and 23s R-RNA structures.* In: Hartman, H. & Matsuno, K., eds. *The origin and evolution of prokaryotic and eukaryotic cells*. World Scientific Publ.Co., New Jersey, in press.

24. Shapiro, B.A. (1988) *CABIOS* **4**, 387-397.

25. Shapiro, B.A. & Zhang, K. (1990) *CABIOS* **6**, 309-318.

26. Tai, K. (1979) *J.ACM* **26**, 422-433.

27. Sankoff, D. & Kruskal, J.B., eds. (1983) *Time Warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison.* Addison Wesley, London.

28. Hogeweg, P. & Hesper, B. (1984) *Nucleic Acid Research* **12**, 67-74.

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

30. Konings, D.A.M. & Hogeweg, P. (1989) *J.Mol.Biol.* **207**, 597-614.

31. Wright, S. (1932) *The roles of mutation, inbreeding, crossbreeding and selection in evolution.* In: *Proceedings of the Sixth International Congress on Genetics.*Vol.**1**, pp.356-366.

32. Kauffman, S. & Levin, S. (1987) *J.theor.Biol.* **128**, 11-45.

33. Kauffman, S.A. (1989) *Adaptation on Rugged Fitness Landscapes.* In: Stein, D., ed. *Complex Systems.*SFI Studies in the Sciences of Complexity. Addison Wesley, Redwood City (Cal.), pp.527-618.

34. Kauffman, S.A. & Johnsen, S. (1991) *J.theor.Biol.* **149**, 467-505.

35. Macken, C.A. & Perelson, A.S. (1989) *Proc.Natl.Acad.Sci.USA* **86**, 6191-6195.

36. Weisbuch, G. (1990) *J.theor.Biol.* **143**, 507-522.

37. Fontana, W., Griesmacher, T., Schnabl, W., Stadler, P.F. & Schuster, P. (1991) *Mh.Chem.* **122**, 795-819.

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

39. Avis, D. (1981) *Can.J.Math.* **33**, 795-802.

40. Eigen, M., McCaskill, J. & Schuster, P. (1989) *Adv.Chem.Phys.* **75**, 149-263.

41. Maynard-Smith, J. (1970) *Nature* **225**, 563-564.

42. Perelson, A.S. & Oster, G.F. (1979) *J.Theor.Biol.* **81**, 645-670.

43. Segel, L.A. & Perelson, A.S. (1988) *Computations in Shape Space: a New Approach to Immune Network Theory.* Perelson, A.S., ed. *Theoretical Immunology. Part Two.* Addison Wesley, Redwood City (Cal.), pp.321-343.

44. Huynen, M., Konings, D.A.M. & Hogeweg, P. (1992) *The effect of multiple coding on the evolutionary properties of RNA secondary structure.* Preprint.

45. Weinberger, E. (1990) *Biol.Cybern.* **63**, 325-336.

46. Wakeman, J.A. & Maden, B.E.H. (1989) *Biochem.J.* **258**, 49-56.

47. Ramek, M. (1990) *Chemical structure formulæ and x/y diagrams with TEX.* In: Clark, M., ed. *TEX: Applications, Uses, Methods. Proceedings of the TEX 88 Conference.* Ellis Horwood Publ., Chichester (U.K.), pp.227-258..

Contents

1. Introduction 2

2. Statistics of elements of RNA secondary structures 5

3. Comparison with natural sequences 10

4. Distances between structures 15

5. Landscapes, combinatory maps, and density surfaces 19

6. Autocorrelation functions and correlation lengths 24

7. Concluding remarks 26

Acknowledgements 27

References 29