Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration

I. Neutral Networks

By

Walter Grüner[a], Robert Giegerich[b], Dirk Strothmann[b], Christian Reidys[a], Jacqueline Weber[a], Ivo L. Hofacker[c], Peter F. Stadler[c,d], and Peter Schuster[a,c,d,*]

[a]Institut für Molekulare Biotechnologie
Beutenbergstraße 11, D-07708 Jena, Germany

[b]Technische Fakultät, Univ. Bielefeld
Postfach 100131, D-33501 Bielefeld, Germany

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

[d]Santa Fe Institute
1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Monatshefte für Chemie, in press

*Mailing Address:
Institut für Molekulare Biotechnologie
Beutenbergstraße 11, PF 100 813, D-07708 Jena, Germany
Phone: **49 (3641) 65 6444 Fax: **49 (3641) 65 6446
E-Mail: pks@imb-jena.de

Abstract

Global relations between RNA sequences and secondary structures are understood as mappings from sequence space into shape space. These mappings are investigated by exhaustive folding of all **GC** and **AU** sequences with chain lengths up to 30. The computed structural data are evaluated through exhaustive enumeration and used as an exact reference for testing analytical results derived from mathematical models and sampling based on statistical methods. Several new concepts of RNA sequence to secondary structure mappings are investigated, among them that of *neutral networks* (being sets of sequences folding into the same structure). Exhaustive enumeration allows to test several previously suggested relations: the number of (minimum free energy) secondary structures as a function of the chain length as well as the frequency distribution of structures at constant chain length (commonly resulting in generalized forms of Zipf's law).

1. Introduction

Conventional biophysics considers sequence structure relations of biopolymers primarily with respect to the folding problem: given is a sequence; which structure does it form under the specified experimental conditions? Such a condition is, for example, the thermodynamic equilibrium for minimum free energy structures. Kinetically determined structures correspond to the outcome of the folding process under certain conditions. Many problems in current molecular biology and biotechnology [1], however, raise questions that cannot be answered satisfactorily by this approach. Required is instead a different view that considers the set of all (possible) sequences as an entity which is mapped onto the set of all (possible) structures. Such problems are, for example, the sensitivity of structures against mutations in the underlying sequences [2], the influence of nucleotide distributions (%**A**, %**U**, %**G**, %**C**) on structures [2, 3] as well as the inverse folding problem: given is a RNA secondary structure; which sequences do fold into this structure under the specified conditions [4, 5]?

The evolution of RNA molecules in replication assays, viroids and RNA viruses can be viewed as an adaptation process on a fitness landscape in the sense of Sewall Wright's imagination [6]. The dynamics of evolution is thus tightly linked to the structure of an underlying landscape. Global features of landscapes can be described by statistical measures like numbers of optima, lengths of walks, and correlation functions, see, for example, [7, 8, 3, 9]. Statistical characteristics of RNA landscapes are accessible on the level of secondary structures by mathematical analysis and computer calculations: these RNA landscapes belong to the same class as well known optimization problems and simple spin glass models [5, 10].

The notion of a landscape has been extended to combinatory maps, thereby allowing for a direct statistical investigation of the sequence structure relationships of RNA at the level of secondary structures [2, 5]. Extensive computational studies have revealed that the frequencies of structures are highly non-uniform, that sequences sharing the same structure are distributed randomly over sequence space, that there exist neutral paths in sequence space along which structures remain uneffected by mutations, and that any desired secondary structure is formed by sequences that can be found close to an arbitrary initial sequence. These results provide convincing evidence that RNA landscapes are as simple as they could possibly be for evolutionary adaptation. The consequences for evolutionary optimization, the early stages of life, and molecular biotechnology are immediate. Based on these findings a random-graph theory was developed [11] that explains the structure of neutral networks in terms of a single parameter: the frequency of neutral mutations. The predictions of this theory, among them connectedness and density of the neutral networks cannot be verified by a statistical approach based on sampling tiny fractions of sequence space. Comparison of the results derived from the random graph approach with real RNA folding data allows to separate generic properties of sequence to structure mappings from nucleotide specific biochemical phenomena.

In this contribution we report the computational techniques that are necessary to exhaustively generate *all* sequences and their secondary structures for two-letter alphabets up to a chain length of $n = 30$. In essence, the main task is easily stated: compute the secondary structures of all sequences and then group together the sequences that fold into the same secondary structure, i.e., produce an explicit representation of *all* neutral networks. Once this has been done the analysis proceeds by determining the geometric structure of these networks.

This paper is organized as follows: In section 2 we review the folding algorithm and the data structures used to represent RNA structures. In section 3 we present data obtained directly from the exhaustive search, such as the over-all number of different minimum free energy structures, the fraction of open structures, and the distribution of preimage sizes. Criteria will be derived that allow to distinguish common and rare structures.

In a forthcoming paper [12] we shall analyze and discuss the internal structures of neutral networks, in particular the size distributions of their connected components. In addition relative locations of neutral networks will be discussed and strong evidence will be presented for *shape space covering*: almost all common structures can be found within a fairly small ball around any random sequence. The results are particularly valuable for a comparison of real data with the results of the random graph approach.

2. RNA Secondary Structures

2.1. RNA Folding

The biochemical and biophysical properties of RNA molecules are determined by their spatial structures. In case of RNA the process of folding the one-dimensional primary structure (sequence) into the three-dimensional tertiary structure can be decomposed into two steps:

(1) Folding of the sequence into a secondary structure by formation of complementary Watson-Crick base pairs, $\mathbf{G}\equiv\mathbf{C}$ and $\mathbf{A}=\mathbf{U}$, and the weaker $\mathbf{G}-\mathbf{U}$ pairs.

(2) Formation of the three-dimensional tertiary structure from the planar pattern.

Such a decomposition is meaningful since the intramolecular forces stabilizing the secondary structures — base pairing and base pair stacking — are much stronger than those accounting for arrangement of the secondary structure elements in space. Thus the free energy of formation for the three-dimensional structure can be estimated by the free energy of the formation for the secondary structure. The dominant role of secondary structures is also well documented in nature since secondary structure elements are conserved in evolution [13, 14, 15, 16].

Table 1. Folding Algorithms for RNA Secondary Structures.

Algorithm	ψ	Abbr.	Remark	Reference
deterministic				
Minimum Free Energy	$-$	`MFE`	fast	[17, 18]
Kinetic Folding	$+$	`KIN`	fast	[19]
5'-3' Folding	$+$	`5-3`	fast	[20]
Partition Function	$-$	`PF`	ensemble	[21]
Maximum Matching	$-$	`MM`	unrealistic	[22]
stochastic				
Simulated Annealing	$+$	`SA`	very slow	[23, 24]

$^{\psi}$ Pseudo-knots can be included. The major problem with the prediction of pseudo-knots is, however, the lack of sufficient experimental energy parameters.

A variety of computer programs predicting RNA secondary structures have been published. A very brief overview is given in Table 1. Two public domain packages for RNA folding are currently available by anonymous `ftp`: Zuker's `mfold` [25] and the `Vienna RNA Package` [26]. All these programs make use of essentially the same energy model for the formation of secondary structures. It explicitly assumes that there are no knots or pseudo-knots.

Each secondary structure is viewed as being composed of stacked base pairs, loops, and external elements which are neither part of a stack nor of a loop. For the sake of a uniform notation two stacked base pairs can be viewed formally as a special type of loop consisting of exactly four nucleotides. Depending on the topology of the loop one distinguishes different loop types: *hairpin loops* have only one base pair, *stacked base pairs* consist of exactly two base pairs, *bulges* have two base pairs adjacent to each other and at least one unpaired base, *interior loops* have two base

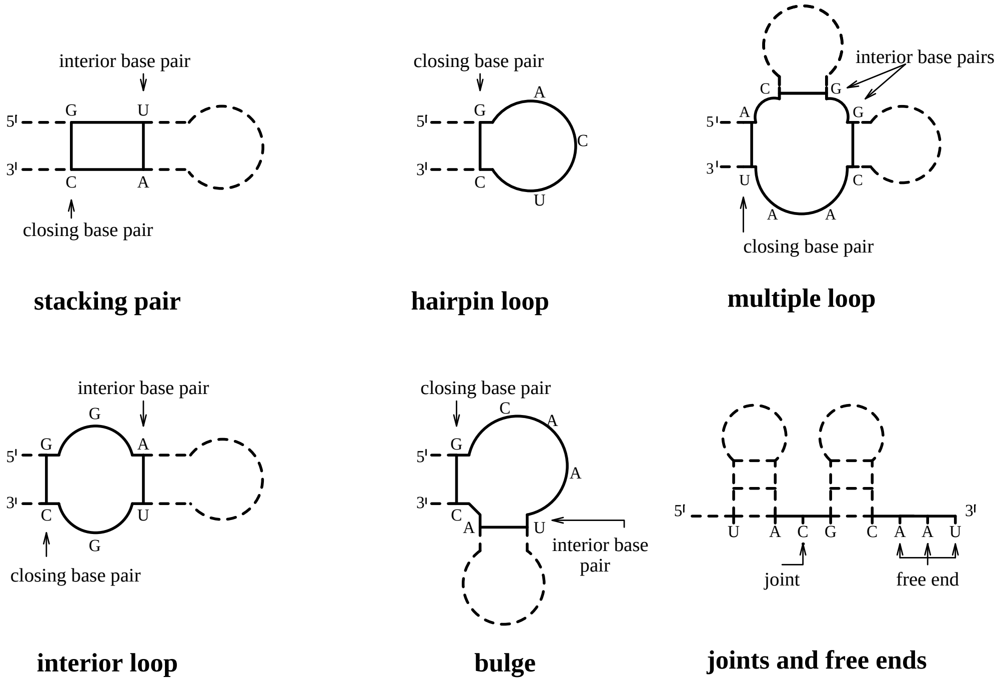

Figure 1: RNA secondary structures can be decomposed into loops and external elements. The energy of a particular structure is the sum of energy contributions for all its loops.

pairs which are not adjacent to each other, and *multi-loops* contain at least three base pairs, see Figure 1. The energy of a secondary structure is the sum of energy contributions of all loops in the structure. These contributions depend on the loop type, the loop size, and partly on the particular sequence of nucleotides in the loop. The individual energy parameters were determined experimentally, see e.g. [27, 28, 29]. While the prediction of the folding algorithm for a particular sequence depends strongly on the details of the parameter set, such details do not affect the global features of the sequence structure relations [30].

The data used in this contribution have been produced with the `fold()` algorithm contained in the `Vienna RNA Package` [26]. It is based on a high performance implementation [31] of the Zuker-Sankoff algorithm [17]. The energy parameters used in this package are an updated version of [29] provided by Danielle Konings [32].

2.2. Representation of Secondary Structures

RNA secondary structures are commonly drawn as *secondary structure graphs* in the biochemical literature. Equivalently, though less intuitively, *circle plots* are used sometimes: therein the sequence is arranged in a circle, and base pairs are indicated as chords connecting the pairing nucleotides. The non (pseudo)knot condition implies that chords do not intersect in this representation. A number of equivalent representations have been developed for special applications. Secondary structures can be translated into rooted planar trees by mapping base pairs to internal nodes and unpaired nucleotides to leaves [2] (This *tree representation* is equivalent to a *mountain representation* [33] obtained by indicating a nucleotide which is paired with a base towards the 3' end by a positive slope, a base with pairing partner to the 5' end by a negative slope, and an unpaired base by horizontal line segment). Both the tree representation and the mountain representation have been used for comparing secondary structures and for computing similarity measures between different secondary structures [15, 34].

The tree and the mountain representation are equivalent to a string representation. A base paired with a partner towards the 3'-end is denoted by '(', a nucleotide pairing with a partner towards the 5'-end by ')', and '.' is used for representing unpaired positions. Each secondary structure is then uniquely determined by a string of length n taken from the alphabet '.()'. Not all strings formed from this alphabet are valid secondary structures: biophysical constraints require that a hairpin loop contains at least three unpaired bases and that each open bracket '(' has to be matched by a closed bracket ')'. It is easy to see that the set of all valid secondary structures of chain length $n \geq 3$ is in fact generated by the context free grammar

$$G : S \rightarrow \quad \text{`...'} \quad | \quad \text{`.'}S \quad | \quad S\text{`.'} \quad | \quad SS \quad | \quad \text{`('}S\text{`)'} \; .$$

This representation is used for I/O in the `Vienna RNA Package`. It is also used in this contribution. Given a string like '((...).)', its parse according to the grammar G uniquely determines pairs of matching parentheses. Similar grammars are used for modeling tRNA structure families in [35, 36]. Relating some sequence x to some structure s, we call x_i, x_j to be in *contact* or *matched '('-and ')'-positions* when the positions i and j are a base-pair in s, i.e. they correspond to matching parentheses in s.

2.3. Folding as a Combinatory Map

Natural RNA sequences are strings of length n over the alphabet $\{\mathbf{G,C,A,U}\}$. The canonical distance measure between two sequences of equal length n is the Hamming distance [37] counting the positions in which to end-to-end aligned sequences differ.

The relation between sequences and structures is mediated by a folding algorithm, see above, in our case the routine `fold()` of the `Vienna RNA Package`. In the reminder of this paper we will write $f(x)$ instead of $\texttt{fold}(x)$. The biological question of sequence-structure relations translates into the mathematical question of determining properties of the mapping

$$f:\, \mathcal{Q}^n_\alpha \to \mathcal{Y}_n\,,$$

where $\mathcal{Q}^n_\alpha$ is the generalized hypercube (Hamming graph) of dimension n over an alphabet of size α and $\mathcal{Y}_n$ is the *shape space* of all secondary structures of length n. The shape space can be viewed as metric space by either using the trivial metric or one of the distance measures based on the tree representation or mountain representation of secondary structures.

The term *combinatory map* was introduced for mappings from a discrete configuration space into some metric space as a generalization of the notion of a *landscape* [9]. It is clear that the folding map f is not necessarily injective, since there are less than 3^n secondary structures as compared to 4^n sequences. From the structure statistics given in [2] immediately follows that f is not onto in general: structures, as derived from the grammar G, typically contain many isolated (non-stacked) base pairs. Consequently, the preimage $f^{-1}(s)$, i.e., the set of all sequences actually folding into s, will be in general a fairly large set. The structure statistics [2] indicates, however, that only a vanishingly small fraction of all secondary structures obtained from folding contains isolated base pairs. Consequently, the image of the sequence space, $f(\mathcal{Q}^n_\alpha)$, will in general be a proper subset of $\mathcal{Y}_n$.

The obvious questions to ask about f are thus the following

- How large is the image of sequence space $f(\mathcal{Q}^n_\alpha)$?
- How large are the preimages of given structures?
- What is the distribution of preimage sizes?
- How many sequences do not form secondary structures, i.e., how large is the preimage of the open structure?
- How is $f^{-1}(s)$ embedded in sequence space?

Many of these questions have been at least partially answered in previous papers describing non-exhaustive computer simulations [3, 9, 2, 5, 10, 4], as mentioned in the introduction. Based on these results a random graph model [11] was conceived that allows to construct the preimages of a (given) secondary structures s as functions of a single parameter, the (average) fraction of neutral neighbors of the sequences folding into s (see the section on neutral networks below).

The random graph model reveals the generic properties of sequence to structure mappings based on base pairing. It is based merely on the existence of some definitions of legal pairings that need not be the complementary pairs (preferential **AA** or **GG** pairs as observed in homo-DNA [38] would be equally acceptable). The exhaustive enumeration data presented here can be used to check the predictions of the random graph model and they may help to detect and interpret systematic deviations of data obtained by RNA folding from this idealized reference.

2.4. Compatibility of Sequences with Structures

The mere notion of base pairing implies an *a priori* relation between sequences and structures which depends only on the legal base pairs in a given alphabet, but which is completely independent of the particular values of the energy parameters. For instance, the legal base pairs in the natural **AUGC** alphabet are **AU**, **UA**, **GC**, **CG**, **GU**, and **UG**. Note that we neglect non-standard pairings such as **AA**, **GA**, **AG**, or **UU** which have been observed in natural RNA structures [39, 40, 41, 42].

Definition. A sequence x is compatible with the secondary structure s if (x_i, x_j) is a legal base pair for all matched '('- and ')'-positions i and j [11]. Given a secondary structure s we shall denote the set of all compatible sequences by $\mathbf{C}[s]$.

Of course, the mere fact that $x \in \mathbf{C}[s]$ does by no means imply that $f(x) = s$, i.e., that x really folds into the secondary structure s. Trivially, we have however $f^{-1}(s) \subseteq \mathbf{C}[s]$, i.e., all sequences folding into a given secondary structure have to be compatible with it. The notion of compatible sequences is essential for any deeper understanding of sequence structure relationships in RNA.

2.5. Neutral Networks

2.5.1. Rearranging the set of compatible sequences

Since our main interest is the representation of $f^{-1}(s)$ in sequence space let us first consider the geometry of the sets $\mathbf{C}[s]$ of compatible sequences in the sequence space $\mathcal{Q}^n_\alpha$. The structure of $\mathbf{C}[s]$ is complicated by the difference between paired and unpaired positions. $\mathbf{C}[s]$ does not in general give rise to a connected subgraph of sequence space, but decomposes into hyperplanes which are characterized by a particular choice of the base pairs at paired positions. "Neighboring" hyperplanes have either Hamming distance $d_H = 1$, for instance if **GC** is replaced by **GU**, or Hamming distance $d_H = 2$ if **GC** is replaced by any other legal pairing (**AU**, **UA**,

CG, **UG**). Note that **GC**→**GU**→**AU** gives rise to three hyperplanes which are connected with each other!

Allowing for both point mutations and base pair exchanges redefines the neighborhood relations in $\mathbf{C}[s]$ by introducing new edges. We shall denote this graph by $\mathcal{C}[s]$. It is clearly connected. In the following we will examine its structure in detail. Given a seconary structure s, a sequence $x \in \mathcal{C}[s]$ has any one of the α letters at each position i for which $i \leftrightarrow$ '.', while positions j and k corresponding to matching brackets '(' and ')' are positions that are occupied by any one of the β legal base pairs. Thus a sequence $x \in \mathcal{C}[s]$ can be represented by the nucleotides in its unpaired positions, and the encoding base pairs in terms of letters from an alphabet of size β at the positions of the open brackets. All letters at the closed brackets can then be deleted. An example is given in Figure 2.

```
.((((((...)))...))).    .((((((...)))...))).    structure
CGCCGGCGGCGCCCGGGGCC    AUGGGUCUCCGACAGUCCGG    sequence
C011001GGC___CGG___C    A500031UCC___AGU___G    base pairs reduced
1011001001___100___1    2500031311___203___0    unpaired bases reduced
10110010011001          25000313112030          reduced sequence


 G=0   C=1   A=2   U=3
GC=0  CG=1  AU=2  UA=3  GU=4  UG=5
```

Figure 2: The effect of procedure `reduce`.

Later in this paper we will make use of the procedure `reduce` performing this contraction of the sequence:

$$\begin{aligned}
\texttt{reduce}(s,' ') &= ' ' \quad , \\
\texttt{reduce}(s,'.'x) &= '.'\texttt{reduce}(s,x) \quad , \\
\texttt{reduce}(s,'('x) &= '('\texttt{reduce}(s,x) \quad \text{and} \\
\texttt{reduce}(s,')'x) &= \texttt{reduce}(s,x) \quad .
\end{aligned}$$

Given the reduced sequence **and** the structure, one can of course reconstruct the original sequence. In fact, it is easy to write a function `expand` such that

$$\texttt{expand}(s, \texttt{reduce}(s, x)) = x$$

for all $s \in \mathcal{Y}$ and all $x \in \mathbf{C}[s]$.

Reduced strings can be rearranged further: we first write the u unpaired positions and then the p base pairs ($n = u + 2p$). Thus we can interpret $\texttt{reduce}(s, x)$ as an element of the direct product space $\mathcal{Q}_\alpha^u \times \mathcal{Q}_\beta^p$, i.e., $\mathcal{C}[s]$ is isomorphic to the product of two Hamming graphs with possible distinct alphabets, which correspond to the unpaired and paired positions, respectively. As an immediate consequence of these considerations we note the following

Lemma. Let $x, y \in \mathbf{C}[s]$. The graphical distance of x and y in the graph $\mathcal{C}[s]$ coincides with the Hamming distance of their reduced representations:

$$d_{\mathcal{C}}(x, y) = d_H(\,\texttt{reduce}(s, x), \texttt{reduce}(s, y)\,)\,.$$

In particular, there is an edge in the graph $\mathcal{C}[s]$ if and only if

$$d_H(\,\texttt{reduce}(s, x), \texttt{reduce}(s, y)\,) = 1\,.$$

Using reduced sequences provides two advantages:

(1) The storage requirements are reduced. We can represent the preimage $f^{-1}(s)$ now as the pair (s, X), where X is the list of all *reduced* sequences folding into s. Note that the length of a reduced sequence is

$$|\texttt{reduce}(s, x)| = u + p = n - p.$$

In the case of the **GC** alphabet this reduces the memory requirements by about 25%.

(2) Comparison of two sequences is more efficient, again because the representation is shorter.

(3) The distance between two sequences can be computed very efficiently since it coincides with their Hamming distance by the above lemma.

2.5.2. Definition of neutral networks

Mutations which do not affect the fitness of an organism are called *neutral* in biology. By the same token, the term neutral is used in the context of RNA to mean mutations which do not alter the (secondary) structure. Hence $f^{-1}(s)$ contains all the sequences which can be considered as neutral mutants of each other.

Definition. The set $f^{-1}(s)$ considered as an induced subgraph of $\mathcal{C}[s]$ is called the *neutral network*, $\mathcal{N}(s)$, of the secondary structure s.

The motivation for the term "network" will become clear later on. Defining $\mathcal{N}(s)$ as induced subgraph of $\mathcal{C}[s]$, rather than as induced subgraph of the sequence space Q^n_α itself, avoids the peculiarities introduced by the logic of base pairing. On the other hand, the neighborhood relation no longer coincides with the action of mutation. Hence we have traded technical tractability for biophysical interpretation.

The most important characteristic of a neutral network $\mathcal{N}(s)$ is its connectivity. In order to retain as much biologically relevant information as possible we consider the unpaired and the paired part of the sequence separately and define $\lambda_u(s)$ and $\lambda_p(s)$ as the average fraction of neutral mutations in Q^u_α and Q^p_β, respectively.

2.5.3. Neutral networks of RNA secondary structures

Let us briefly summarize the properties of the neutral networks of RNA minimum free energy structures.

- The distribution of preimage sizes follows roughly a so called generalized Zipf's law, i.e., their rank-order statistics follows a distribution function of the form

  $$\phi(r) = A(1 + r/B)^{-\gamma}$$

  where r is the rank-order of a structure, $\phi(r)$ its frequency, A gives the abundance of the most frequent structures, $B > 0$ measures the number of "frequent" structures, and $\gamma > 1$ determines the shape of the power-law tail [5].

- A *neutral path* is a path $\{x_0, x_1, \ldots, x_\ell\}$ in $\mathcal{N}(s)$ such that $d_{\mathcal{C}}(x_0, x_\ell) = \ell$, i.e., a neutral path if obtained by selecting a start sequence x_0 folding into s and then successively choosing neutral neighbors in such a way that the (graphical) distance to the starting sequence is increased with each accepted step. The length ℓ of a neutral path is therefore a lower bound on the diameter of the neutral network $\mathcal{N}(s)$. Computer studies based on RNA folding have shown that the average values of ℓ are much larger than the distance between randomly chosen sequences for all common structures. A precise definition of a common structure will be given below, see sect. 3.2. In this sense, the neutral networks of common sequences reach through all of sequence space, or more precisely, the set of compatible sequences. [5, 30].
- Inverse folding provides a means of estimating the distance from a random sequence to the nearest sequence folding into a desired target structure s. One finds that sequence space is covered with fairly small balls each of which contains almost all common structures. The radius R of these balls is only slightly larger than the average distance from a random sequence to the set of compatible sequences for any common structure s. This property has been termed shape space covering in [5].
- The connectivities λ_p and λ_u become constant for large n when averaged over large samples of randomly chosen sequences [11], i.e., RNA sequence-structure maps are characterized by a very high degree of neutrality.

2.5.4. *Random graph theory of neutral networks*

The findings from the computational studies as listed above prompted us to search for a generic statistical model (with as few parameters as possible) that could explain the data. As a first step we have recently proposed a random graph model explaining structures and properties of individual neutral networks [11].

As already stated $\mathcal{C}[s]$ is the direct product of the hypercubes $\mathcal{Q}^u_\alpha \times \mathcal{Q}^p_\beta$ and proceed by constructing random subgraphs in each of the graphs $\mathcal{Q}^u_\alpha$, $\mathcal{Q}^p_\beta$ separately. This is done by selecting each vertex of $\mathcal{Q}_\alpha^u$, $\mathcal{Q}_\beta^p$ with independent probabilities λ_u and λ_p. Here "vertex in $\mathcal{Q}_\alpha^u$" means the "unpaired" part of the sequence x and "vertex in $\mathcal{Q}_\beta^p$" means the "paired" part of the sequence x (that is compatible wih s). Thereby one obtains *randomly induced subgraphs* $\Gamma_u < \mathcal{Q}_\alpha^u, \Gamma_p < \mathcal{Q}_\beta^p$ and the neutral network $\mathcal{N}[s]$ is given by

$$\mathcal{N}[s] \stackrel{\text{def}}{=\!=} \Gamma_u \times \Gamma_p \,.$$

The random graph approach does not deal with specific biochemical or biophysical features of the folding process by doing two random selections, one for the unpaired part of a sequence and the other for the paired part. A sequence folds into s with probability $\lambda_u \times \lambda_p$. However, the basic parameters have a biochemical interpretation namely to be for the "unpaired" part the expected fraction of neutral sequences in Hamming distance one and for the "paired" part the fraction of neutral neighbors for simultaneous base pair exchanges. In other words the parameters reflect the stability of the structure s under point mutations in the "unpaired region" and base pair exchanges (i.e., mutations that preserve compatibility) in the "paired regions".

Table 2. Asymptotic values for the fractions of neutral neighbors λ_u and λ_p.

α	*Model I*		*Model II*	Alphabet	*RNA Structures*	
	λ_d^*	λ_c^*	$\lambda_d^* = \lambda_c^*$		unpaired	paired
2	0.5	0.2929	0.5	**GC**	0.271	0.436
				AU	0.352	0.495
4	0.3700	0.2063	0.3700	**GCXK**	0.479	0.509
				GCAU	0.495	
6	0.3011	0.1640	0.3011	**GCAU**		0.455

The theory of randomly induced subgraphs of sequence spaces predicts that, in the limit of long sequences, there exists a critical value λ^* such that a neutral network is a dense and connected subgraph of $\mathcal{C}[s]$ if $\lambda > \lambda^*$. A subgraph G of a finite graph H ($G < H$) is dense in H if and only if each vertex of H is either in G or has at least a neighbor in G. Neutral networks show a typical percolation phenomenon in sequence space. Moreover for all $\lambda_u, \lambda_p > 0$ there exists in the limit of long sequences a so called *giant component* i. e. most of the sequences are pairwise connected in $\mathcal{C}[s]$. Conversely, the network is (in its projections $\mathcal{Q}^u_\alpha, \mathcal{Q}^p_\beta$) neither dense nor connected if its λ-value is below the critical value. A later and refined version takes into account the different connectivities for unpaired and paired positions. Table 2 compiles the critical connectivities for the two variants of the random graph model as well as for sequences from different alphabets. The predictions of the random graph model concerning the connectivity or percolation problem and the density of subgraphs were checked by exhaustive ennumeration. The data shown in table 2 indicate in full agreement with previous studies [1] that neutral networks of structures **GC**-sequences are more likely to fall below the percolation threshold than those derived from **AU**- or **AUGC**-sequences.

3. Exhaustive Enumeration of Secondary Structures

3.1. Folded Secondary Structures

3.1.1. An upper bound to the number of folded structures

Of course, the number of folded secondary structures cannot exceed the number S_n of possible secondary structures. S_n can be calculated from the simple recursion

$$S_n = S_{n-1} + \sum_{k=m}^{n-2} S_k \, S_{n-k-2} \quad n \geq m+1 \qquad S_0 = S_1 = \ldots = S_{m+1} = 1$$

where m is the minimum number of unpaired digits and which follows directly from the grammar G, see [43]. In the biophysically relevant case $m = 3$ the asymptotics of S_n are given by:

$$S_n^{[1]} \sim 0.7131 \times n^{-3/2}\,(2.2888)^n$$

As mentioned above, isolated base pairs are extremely rare in folded secondary structures. Hence, a better estimate can be obtained by counting the number of secondary structures $S_n^{[2]}$ which do not contain isolated base pairs, i.e., all base pairs are contained in stems of length at least two. A recursion and an asymptotic expression for this series was recently derived [44]. It turns out that $S_n^{[2]}$ is significantly smaller than 2^n:

$$S_n^{[2]} \sim 1.4848 \times n^{-3/2}\,(1.8488)^n$$

Most structures counted this way still exhibit much shorter helices than average folded structures. We expect therefore that $S_n^{[2]}$ still overestimates the number of folded secondary structures.

3.1.2. A lower bound to the number of folded structures

A non-trivial lower bound can be obtained from estimating the size of the set of sequences compatible to a given structure s. Indeed, we know that 50% of all sequences fold into structures with at least $p_0 n$ base pairs, where p_0 is some constant with $0 < p_0 < 1$ independent of n, see [2]. Therefore, 50% of the α^n sequences fold into structures which have compatible sequences of size at most

$$|\mathbf{C}[s]| \leq \alpha^{(1-2p_0)}\beta^{p_0 n} = \alpha^n \cdot \left(\frac{\beta}{\alpha^2}\right)^{p_0 n}$$

where $\beta \leq \alpha^2$ is the number of different legal base pairs in the given alphabet.

Even if all sequences in $\mathbf{C}[s]$ would fold into the same structure, and if all compatible sequences were disjoint[1] there must be at least

$$S_n^{\mathrm{l.b.}} = \frac{1}{2}\left(\frac{\alpha^2}{\beta}\right)^{p_0 n}$$

[1] It was in fact shown that the opposite is true [11]: The compatible sets of any two secondary structures have a non-empty intersection.

Table 3. Lower Bounds on the Number of Folded Structures.

Alphabet	α	β	p_0	$r = (\alpha^2/\beta)^{p_0}$
GC	2	2	0.403	1.322
AU	2	2	0.403	1.278
AUGC	4	6	0.290	1.329
AUGC$^+$	4	4	0.207	1.332
GCXK^a	4	4	0.270	1.453
ABCDEF^a	6	6	0.185	1.393

$^+$ GU pairs suppressed.
^{a} These artificial alphabets contain two (three) complementary base pairs with identical energy parameters.

different secondary structures. Numerical estimates for p_0 can be obtained from the data for the average number of base pairs in folded secondary structures [9], see table 3. The number of base pairs is concentrated around $p_0 n$ for large n. Even if mean and median would differ significantly in this distribution, it would only lead to a (minor) correction of the pre-factor 1/2 without affecting the exponential part of the bound $S_n^{\mathrm{l.b.}}$. It is interesting to note that $r = (\alpha^2/\beta)^{p_0}$ does not strongly depend on the alphabet. Just as the upper bound discussed in the previous subsection, this lower bound is far from being sharp as well.

3.1.3. Exhaustive structure generation

Because of the intractably larger numbers of sequences over alphabets with more than $\alpha = 2$ letters, we restrict our numerical investigations to two letter alphabets. We choose two examples **GC** and **AU**. In the latter case the data are expected to reflect the short-sequence effects much stronger because the energy parameters of **AU** base pairs are such that sequences of tractable length are very likely to be unfolded.

The sequences are generated in collections of five million sequences and their respective complements each. The sequences are folded using `fold()` and then grouped according to their structures. After a section is completed, all sequences

x folding into the secondary structures are appended to the `UNIX` file `FILE`(s). Sequences are stored in 32-bit machine words, with 0 and 1 replacing **G** and **C**, respectively. This method of storage, which limits the sequence length to 32, does not restrict the approach in praxis, since a length of $n = 32$ is already above the limits of both the accessible CPU resources and accessible storage capacities. In fact, longest RNA molecules we have investigated have chain length $n = 30$. This computation required about 130 days of CPU time on an `IBM RISC 6000` workstation with 256 MB RAM, and more than 4.3 GB of disc space.

The bounds on the number of secondary structures were important for organizing the disc storage of the data. In the worst case, the upper bound given above leads us to expect about a million different secondary structures for $n = 30$. This is by far more than the number of files that can be contained in a single directory of the `UNIX` file systems on our workstations. In order to overcome this restriction, the files are arranged in a hierarchical directory structure limiting the number of files per subdirectory to about 1000. For $n = 30$, this directory structure alone occupies about 15.6 MB of disc space.

3.2. Numerical Estimates From Exhaustive Enumeration

3.2.1. The number of minimum free energy structures

The upper and lower bounds on the number of secondary structures, which we have discussed in the previous sections have been based on fairly crude estimates. Much more accurate estimates for the actual number of different structures that are realized by a particular folding algorithm can be obtained by extrapolation from exhaustive counts for short chains.

Linear regressions of a $\log S_n^{\mathrm{MFE}}$ versus n plot of the data shown in figure 3 yields the following estimates

$$\begin{aligned} S_n^{\mathrm{MFE}} &\sim (0.0853 \pm 0.0009) \times (1.6360 \pm 0.0007)^n \quad \text{for } \mathbf{GC}, \\ S_n^{\mathrm{MFE}} &\sim (0.0097 \pm 0.0038) \times (1.489 \pm 0.029)^n \quad \text{for } \mathbf{AU}, \end{aligned}$$

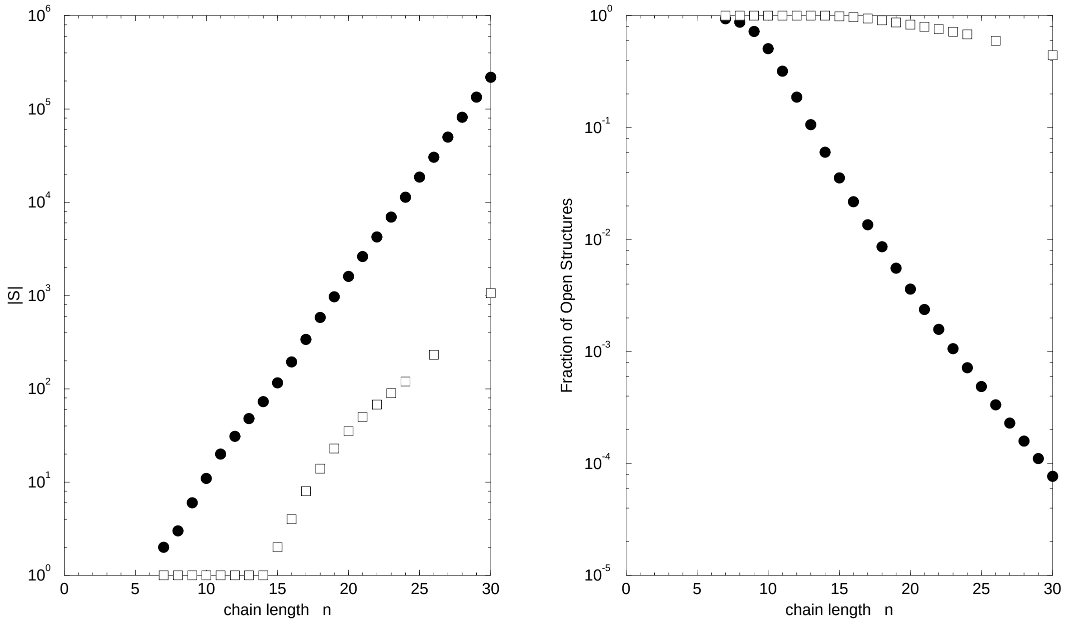

Figure 3: Numbers of structures and abundance of the open structure obtained by exhaustive enumeration computed for **GC**, •, and **AU**, □, alphabets. Numbers of structures $|S|$ are presented as functions of the chain length (l.h.s.). The fractions of sequences that fold in to the open structure are shown on the r.h.s.

The estimate is quite good for the **GC** alphabet, while for the **AU** alphabet finite size effects seems to dominate. This is due to the weaker base pairs, which leave a large fraction of the sequences unfolded in their minimum free energy structure. The exact numbers of different structures can be found in table 4.

The abundance of open structures (of the type '......') is therefore a good indicator for the influence of finite size effects. For **GC** we find a distinct exponential decrease in the fraction of sequences that do not form base pairs, see figure 3. In the **AU** case almost half of the sequences with a chain length of $n = 30$ are shorter do not form base pairs. Consequently we cannot expect reliable estimates for the asymptotics of S_n^{MFE} in this case. We find indeed that number S_n^{MFE} is much smaller for both alphabets then the combinatorial estimate discussed in sect. 3.1.1.

3.2.2. Common structures

Definition. A structure s is said to be *common* if its preimage $f^{-1}(s)$ is not smaller than the average size of a neutral network [1], that is in mathematical notation, if

$$|f^{-1}(s)| \geq \sum_{s' \in \mathcal{S}} \frac{1}{|\mathcal{S}|} |f^{-1}(s')| = \frac{\alpha^n}{|\mathcal{S}|}.$$

Furthermore, let r_c denote the rank of the rarest common structure, i.e., the structure with rank r_c is common, but the structure with rank $r_c + 1$ is not common.

The number of sequences that fold into common structures, n_c is listed in Table 4. Two quantities are of particular interest: the fraction $r_c/|\mathcal{S}|$ of common structures within the set $\mathcal{S}$ of all structures that are obtained by folding, figure 4(l.h.s.), and the fraction n_c/α^n of sequences that fold into common structures, see figure 4(r.h.s.).

The fraction of common structures among all minimum free energy structures decreases with chain length. The data for the **GC** alphabet as consistent with an asymptotically exponential decrease of $r_c/|\mathcal{S}|$. The data for the **AU** alphabet exhibit a minimum at $n = 22$. A trend for long sequence, however, cannot be read off the data. In analogy to the results for **GC** sequences and based on the Zipf's law type distribution of preimage sizes we expect, that $r_c/|\mathcal{S}|$ approaches 0 as n becomes large.

The fraction of sequences that fold into common structures slowly increases in the case of **GC** sequences; for $n = 30$ we have $n_c/2^{30} \approx 0.931$, i.e., less than 7% of all sequences fold into non-common structures. For **AU** we find a minimum at $n = 24$ with about 70% of the sequences folding into common structures, and a steep increase of this fraction for longer chains. The data are consistent with the conjecture that asymptotically almost all sequences will fold into common structures.

Table 4. Common Structures and their Preimages.

	GC			AU		
n	$\|\mathcal{S}\|$	r_c	n_c	$\|\mathcal{S}\|$	r_c	n_c
7	2	1	128	1	1	*
8	3	2	240	1	1	*
9	6	2	435	1	1	*
10	11	5	903	1	1	*
11	20	8	1,744	1	1	*
12	31	14	3,633	1	1	*
13	48	23	7,522	1	1	*
14	73	32	14,860	1	1	*
15	116	44	29,212	2	1	32,256
16	195	65	59,193	4	1	63,488
17	340	87	115,516	8	1	123,960
18	582	118	225,161	14	1	238,366
19	973	184	451,339	23	1	456,964
20	1,610	287	903,567	35	1	875,710
21	2,615	462	1,827,313	50	1	1,673,596
22	4,258	753	3,717,115	68	1	3,185,872
23	6,936	1,203	7,548,569	90	3	6,262,203
24	11,348	1,867	15,248,295	120	3	11,836,758
25	18,590	2,870	30,747,664	–	–	–
26	30,501	4,303	61,718,490	232	21	48,789,050
27	49,949	6,373	123,636,916	–	–	–
28	81,748	9,580	247,910,546	–	–	–
29	133,782	14,642	497,599,300	–	–	–
30	218,820	22,719	999,513,710	1064	192	936,240,694

* All sequence fold into the open structure.

3.3. The Distribution of Preimage Sizes

The distribution of preimage sizes has been studied previously for coarse grained structures on sequences up to length 100 by folding samples of several million random sequences [5, 10]. These studies, which access the distribution of preimage sizes in the realm of common structures and the onset of the tail have suggested the distribution follows a generalized Zipf's law [45, 46] which can be parameterized

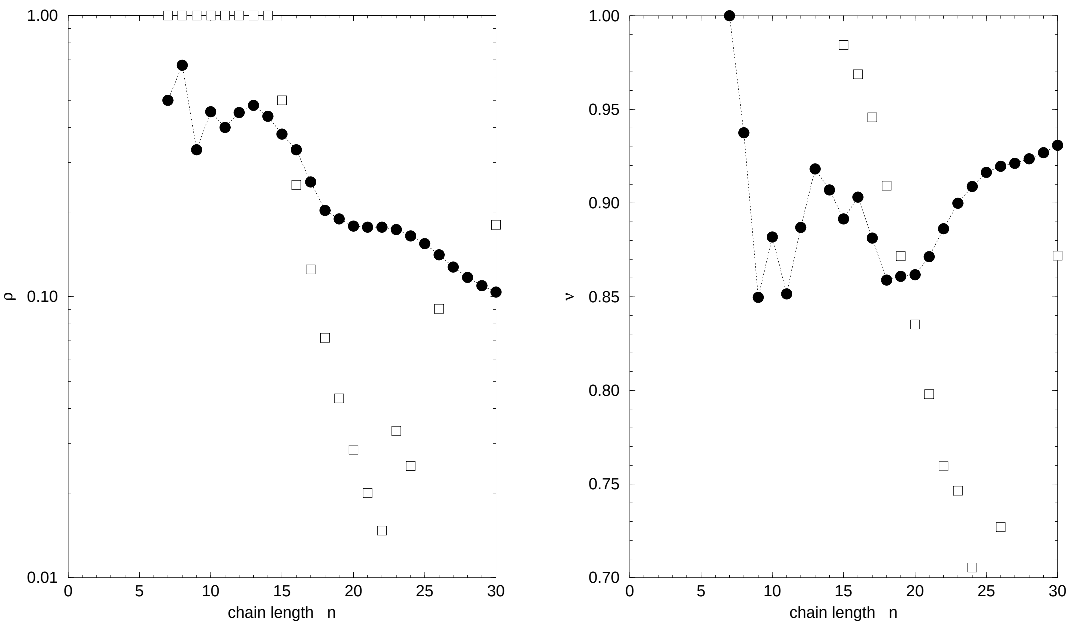

Figure 4: Fraction of common structures and fraction of sequences folding into common structures computed for **GC**, •, and **AU**, □, alphabets. The chain length dependence of $r_c/|\mathcal{S}| = \rho$ (l.h.s.) seems to decrease with increasing chain length whereas the chain length dependence of $n_c/\kappa^n = \nu$ (r.h.s.) strongly indicates convergence towards $\lim_{n\to\infty} \nu = 1$.

as [47]

$$\phi(r) \approx A(1 + r/B)^{-\gamma} \qquad \text{with } B > 0 \text{ and } \gamma > 1,$$

see section 2.5.3.

The exhaustive folding of the entire sequence space allows us to compare this model distribution with the actual rank-order statistics of RNA secondary structures. We find qualitatively similar distributions with few frequent structures and a long tail of very rare structures. Quantitatively, the fit to the model function is rather poor mostly because frequencies should be nearly constant at $r \approx 0$, i.e. the model function exhibits an even sharper transition from common to rare structures see figure 5. This may be partly a finite size effect, since the transition clearly becomes sharper with increasing chain length. Already at chain length 30 more than 93% of **GC** sequences fold into structures with $r < r_c = 22719$, i.e. the most frequent 10% of structures.

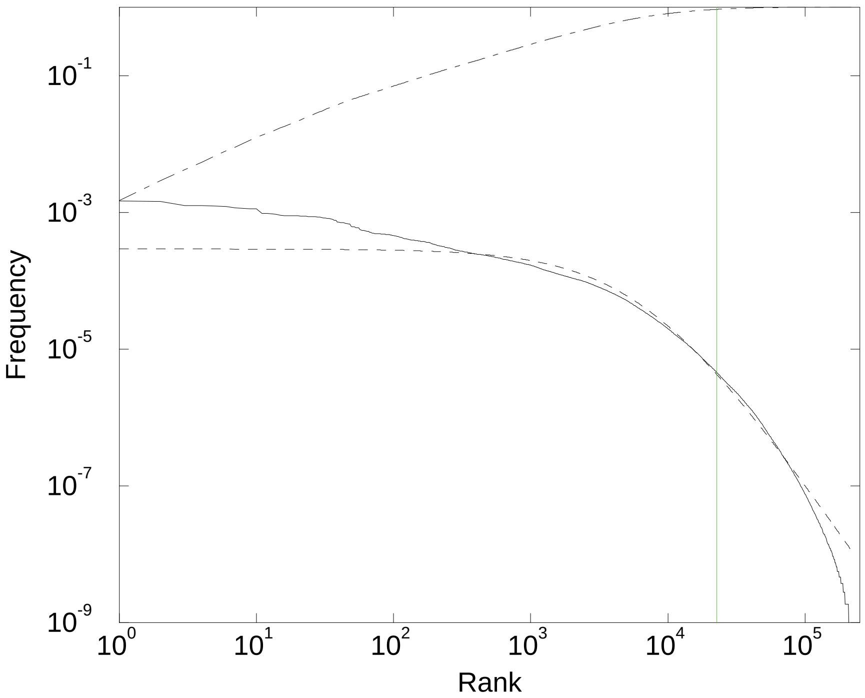

Figure 5: Distribution of preimage sizes for **GC** sequences length 30. The dashed line is a fit to the model above using $A = 0.00029$, $B = 7040$, $\gamma = 2.93$. The dot-dashed line is the integral over the distribution, the vertical line marks $r_c = 22719$.

4. Conclusions

Investigations of complete sets of RNA structures obtained by folding of all sequences of RNA molecules of given chain length n are usually prohibitive because of the hyperastronomically large numbers of sequences. The only known exceptions are the secondary structures formed by short sequences ($n \leq 30$) derived from binary alphabets (**GC** and **AU**). In these cases the numbers of sequences do not appreciably exceed 10^9 and there are less than a quarter of a million different structures. Creation of reference samples by exhaustive folding and retrieval of the desired information, nevertheless, requires special methodologies even in these cases with short chain lengths. Examples of problems that can be addressed by exhaustive enumeration are the elaboration of a clear definition of *common* for RNA secondary structures, the computation of the numbers of (minimum free energy) secondary structures for different base pairing alphabets as functions of the chain length, and the verification of a generalized form of Zipf's law for the frequency distribution of secondary structures at constant chain length.

Different strengths of the interaction between bases (hydrogen bonding and base pair stacking) in the **GC** and **AU** system have a drastic influence on the numbers of minimum free energy structures as well as on the sizes and structures of neutral networks. The effect is twofold:

(1) Weaker interactions in structures built from **AU**-only sequences make finite size effects (e.g. the influence of small chain lengths) more dominant. The fraction of the unfolded or open chain structure, for example, is much larger with **AU**-only than with **GC**-only sequences. In addition, the chain length of the shortest **AU** sequences that form a non-trivial (folded) secondary structure is $n = 15$ whereas **GC** sequences form structures already at $n = 7$. Consequently, the number of different structures formed by **GC** sequences is much larger than the corresponding number for **AU** sequences of the same chain length.

(2) Compared to **GC** stacks more base pairs are needed to form a stable **AU** stack. Hencestacks are longer on the average in structures dominated by **AU**. This has the consequence that the number of structures from **AU** sequences increases less strongly with the chain length ($\propto 1.49^n$) than in the **GC** case ($\propto 1.69^n$).

The definition of *common structures* introduced here is based on simple enumeration: a structure is common if it is formed by more sequences than a fictious average structure whose preimage size is κ^n/S_n sequences. Extrapolation of data computed for different chain lengths suggest a general result that is highly relevant for biotechnology and evolution. The fraction of structures which are common decreases (exponentially) with increasing chain length whereas at the same time the fraction of sequences that fold into the common structures approaches unity. In other words, for long chains a relatively small fraction of all structures is formed by almost all sequences. Most of the structures are thus rare in the sense that they are formed by relatively few sequences only, they will neither be found by natural evolution nor will the play a role in evolutionary biotechnology. The set of common structures thus forms the repertoire from which adaptive processes choose *in vivo* and *in vitro*.

Acknowledgements

The work carried out in Wien has been supported financially by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (Projects 9942-PHY and 10578-MAT). Support of research in Jena by the Commission of the European Communities is gratefully acknowledged (CEC Contract Study PSS*0846). Our research was also continuously supported by the Santa Fe Institute.

References

[1] P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *Journal of Biotechnology*, 41:239–257, 1995.

[2] W. Fontana, D A.M. Konings, P.F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[3] W. Fontana, T. Griesmacher, W. Schnabl, P.F. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degredation rate constants of RNA secondary structures. *Mh.Chem.*, 122:795–819, 1991.

[4] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Mh.Chem.*, 125:167–188, 1994.

[5] P. Schuster, W. Fontana, P.F. Stadler, and I.L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc.Roy.Soc.(London)B*, 255:279–284, 1994.

[6] S. Wright. The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In D. F. Jones, editor, *Int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.

[7] M. Eigen, J. McCaskill, and P. Schuster. The molecular Quasispecies. *Adv. Chem. Phys.*, 75:149 – 263, 1989.

[8] E.D. Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol.Cybern.*, 63:325–336, 1990.

[9] W. Fontana, P.F. Stadler, E.G. Bornberg-Bauer, T. Griesmacher, I.L. Hofacker, M. Tacker, P. Tarazona, E.D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47:2083 – 2099, 1993.

[10] P. Schuster and P.F Stadler. Landscapes: Complex optimization problems and biopolymer structures. *Computers Chem.*, 18:295–314, 1994.

[11] C. Reidys, P.F. Stadler, and P. Schuster. Generic properties of combinatory maps - Neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 1995. Submitted. SFI-Preprint Series No. 95-07-058.

[12] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker, P.F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structure of neutral networks and shape space covering. *Mh.Chem.*, 127:***–***, 1996.

[13] D. Sankoff, A.-M. Morin, and R.J. Cedergren. The evolution of 5S RNA secondary structures. *Can.J.Biochem.*, 56:440–443, 1978.

[14] T.R. Cech. Conserved sequences and structures of groupI introns: Building an active site for RNA catalysis - A review. *Gene*, 73:259–271, 1988.

[15] D.A.M. Konings and P. Hogeweg. Pattern analysis of RNA secondary structure. Similarity and consensus of minimal-energy folding. *J.Mol.Biol.*, 207:597–614, 1989.

[16] S.-Y. Le and M. Zuker. Common structures of the 5' non-coding RNA in enteroviruses and rhinoviruses. Thermodynamical stability and statistical significance. *J.Mol.Biol.*, 216:729–741, 1990.

[17] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull.Math.Biol.*, 46:591–621, 1984.

[18] M. Zuker. The use of dynamic programming algorithms in RNA secondary structure prediction. In Michael S. Waterman, editor, *Mathematical Methods for DNA Sequences*, pages 159–184. CRC Press, 1989.

[19] H.M. Martinez. An RNA folding rule. *Nucl.Acid.Res.*, 12:323–335, 1984.

[20] M. Tacker. *Robust Properties of RNA Secondary Structures.* PhD thesis, Universität Wien, 1993.

[21] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[22] R. Nussinov, G. Piecznik, J.R. Griggs, and D.J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35:68–82, 1978.

[23] A.A. Mironov, L.P. Dyakonova, and A.E. Kister. A kinetic approach to the prediction of RNA secondary structures. *J. of Biomolecular Structure and Dynamics*, 2:953, 1985.

[24] A.A. Mironov and A.E. Kister. RNA secondary structure formation during transcription. *J. of Biomolecular Structure and Dynamics*, 4:1–9, 1986.

[25] M. Zuker. `mfold-2.0`. `ftp://snark.wustl.edu/pub/mfold-sgi-2.2.tar.Z`. (Public Domain Software).

[26] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. `Vienna RNA Package`. `ftp://ftp.itc.univie.ac.at/pub/RNA/ViennaRNA-1.03`. (Public Domain Software).

[27] W. Salser. Globin messenger RNA sequences - analysis of base-pairing and evolutionary implications. *Cold Spring Harbour Symp. Quant. Biol.*, 42:985, 1977.

[28] S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M.H. Caruthers, T. Neilson, and D.H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA*, 83:9373–9377, 1986.

[29] J.A. Jaeger, D.H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc.Natl.Acad.Sci., USA, Biochemistry*, 86:7706–7710, 1989.

[30] M. Tacker, P.F. Stadler, E.G. Bornberg-Bauer, I.L. Hofacker, and P. Schuster. Robust properties of RNA secondary structure folding algorithms. In preparation, 1995.

[31] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophys.Chem.*, 26:123–147, 1987.

[32] D.A.M. Konings. Thermodynamic parameters for RNA folding. Private Communication, 1990.

[33] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucleic acids research*, 12:67–74, 1984.

[34] B.A. Shapiro and K. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *CABIOS*, 6:309–318, 1990.

[35] Y. Sakakibara, M. Brown, R.C. Underwood, I. Saira Mian, and D. Haussler. Stochastic context-free grammars for modeling RNA. Report, UC Santa Cruz, 1993.

[36] Y. Sakakibara, M.Brown, R. Hughey, I.S. Mian, K. Sjölander, R.C. Underwood, and D. Haussler. The Application of Stochastic Context-Free Grammars to Folding, Aligning and Modeling Homologous RNA Sequences. Report, UC Santa Cruz, 1993.

[37] R. W. Hamming. Error detecting and error correcting codes. *Bell Syst. Tech. J.*, 29:147–160, 1950.

[38] A. Eschenmoser. Hexose nucleic acids. *Pure and Applied Chemistry*, 65:1179–1188, 1993.

[39] J. Santa Lucia Jr., R. Kierzek, and D.H. Turner. Effects of GA mismatches on the structure and thermodynamics of RNA internal loops. *Biochemistry*, 29:8813–8819, 1990.

[40] I. Tinoco Jr., M. Chastain, and X. Chen. Non-watson-crick base pairs and rna structure. *Clinical chemistry*, 40:646, 1994.

[41] H.W. Pley, K.M. Flaherty, and D.B. McKay. Three-dimensional structures of a hammerhead ribozyme. *Nature*, 372:68–74, 1994.

[42] H.W. Pley, K.M. Flaherty, and D.B. McKay. Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. *Nature*, 372:111–113, 1994.

[43] M.S. Waterman. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.*, 1:167 – 212, 1978.

[44] I.L. Hofacker, P. Schuster, and P.F. Stadler. Combinatorics of RNA secondary structures. *Submitted to SIAM J. Disc. Math.*, 1993.

[45] G.K. Zipf. *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, Reading, MA, 1949.

[46] Y.S. Chen. Zipf's law in text modeling. *Int. J. General Systems*, 15:232, 1989.

[47] B.B. Mandelbrot. *The Fractal Geometry of Nature.* Freeman & Co., New York, 1983.

Table of Contents

1. Introduction 1

2. RNA Secondary Structures 3

2.1. RNA Folding 3

2.2. Representation of Secondary Structures 5

2.3. Folding as a Combinatory Map 7

2.4. Compatibility of Sequences with Structures 8

2.5. Neutral Networks 9

2.5.1. Rearranging the set of compatible sequences 9

2.5.2. Definition of neutral networks 11

2.5.3. Neutral networks of RNA secondary structures 12

2.5.4. Random graph theory of neutral networks 13

3. Exhaustive Enumeration of Secondary Structures 15

3.1. Folded Secondary Structures 15

3.1.1. An upper bound to the number of folded structures 15

3.1.2. A lower bound to the number of folded structures 16

3.1.3. Exhaustive structure generation 17

3.2. Numerical Estimates From Exhaustive Enumeration 18

3.2.1. The number of minimum free energy structures 18

3.2.2. Common structures 19

3.3. The Distribution of Preimage Sizes 20

4. Conclusions 23

Acknowledgements 25

References 25