Bulletin of Mathematical Biology (1999) **61**, 437–467
Article No. bulm.1998.0085
Available online at http://www.idealibrary.com on IDEAL®

RNA Structures with Pseudo-knots: Graph-theoretical, Combinatorial, and Statistical Properties

CHRISTIAN HASLINGER
Institut für Theoretische Chemie,
Universität Wien
Währingerstraße 17,
A-1090 Wien,
Austria
E-mail: grisu@tbi.univie.ac.at

PETER F. STADLER[*]
The Sante Fe Institute,
1399 Hyde Park Road,
Sante Fe,
NM 87501, U.S.A.
E-mail: studla@tbi.univie.ac.at or stadler@santafe.edu

The secondary structures of nucleic acids form a particularly important class of contact structures. Many important RNA molecules, however, contain pseudo-knots, a structural feature that is excluded explicitly from the conventional definition of secondary structures. We propose here a generalization of secondary structures incorporating 'non-nested' pseudo-knots, which we call *bi-secondary structures*, and discuss measures for the complexity of more general contact structures based on their graph-theoretical properties. Bi-secondary structures are planar trivalent graphs that are characterized by special embedding properties. We derive exact upper bounds on their number (as a function of the chain length n) implying that there are fewer different structures than sequences. Computational results show that the number of bi-secondary structures grows approximately like 2.35^n. Numerical studies based on kinetic folding and a simple extension of the standard energy model show that the global features of the sequence-structure map of RNA do not change when pseudo-knots are introduced into the secondary structure picture. We find a large fraction of neutral mutations and, in particular, networks of sequences that fold into the same shape. These neutral networks percolate through the entire sequence space.

1. Introduction

Presumably the most important problem and the greatest challenge in present day theoretical biophysics is deciphering the code that transforms sequences of

[*]Address for correspondence: P. F. Stadler, Institüt für Theoretische Chemie, Universität Wien, Währingerstr. 17, A-1090 Wien, Austria.

0092-8240/99/030437 + 31 $30.00/0

biopolymers into spatial molecular structures. A sequence is properly visualized as a string of symbols which together with the environment encodes the molecular architecture of the biopolymer. In case of one particular class of biopolymers, the ribonucleic acid (RNA) molecules, decoding of information stored in the sequence can be properly decomposed into two steps: (i) formation of the secondary structure, that is, of the pattern of Watson–Crick (and **GU**) base pairs, and (ii) the embedding of the contact structure in three-dimensional space.

The sequence structure relation of RNA was studied in detail in a series of papers (Fontana *et al*., 1991, 1993a, b; Bonhoeffer *et al*., 1993; Schuster *et al*., 1994; Tacker *et al*., 1994; Grüner *et al*., 1996a,b; Tacker *et al*., 1996) at the level of secondary structures. The most salient findings of these investigations are:

(i) There are many more sequences than (secondary) structures.
(ii) There are few frequent and many rare structures. Almost all sequences fold into frequent or 'common' structures.
(ii) Sequences that fold into a 'common' structure are distributed nearly uniformly in sequence space.
(iv) A sequence folding into a 'common' structure has a large number of neutral neighbors (folding into the same structure) and a large number of neighboring sequences that fold into very different secondary structures.
(v) Neutral paths percolate sequence space along which all sequences fold into the same secondary structure. In fact, there are extended *neutral networks* of sequences folding into the same 'common' structure (Grüner *et al*., 1996b; Reidys and Stadler, 1996).
(vi) Almost all 'common' structures can be found close to any point in sequence space. This property is called *shape space covering*.

The impact of these features on evolutionary dynamics is discussed in Schuster (1995) and Huynen *et al*. (1996): a population explores sequence space in a diffusion-like manner along the neutral network of a viable structure. Along the fringes of the population novel structures are produced by mutation at a constant rate (Huynen, 1996). Fast diffusion together with perpetual innovation makes these landscapes ideal for evolutionary adaptation (Fontana and Schuster, 1998).

The 'classical' definition of secondary structures incorporates a quite restrictive condition on the set of base pairs that implies a tree-like arrangement of the double-helical regions, see Fig. 1. Additional interactions between different branches of this tree are referred to as *pseudo-knots* (for an exact definition see Section 2). Pseudo-knots are excluded from many studies for a mostly technical reason (Waterman and Smith, 1978a, b): the folding problem for RNA can be solved efficiently by dynamic programming (Waterman and Smith, 1978b; Zuker and Sankoff, 1984) in their absence.

On the other hand, an increasing number of experimental findings, as well as results from comparative sequence analysis, suggest that pseudo-knots are important structural elements in many RNA molecules (Westhof and Jaeger, 1992). Notably,

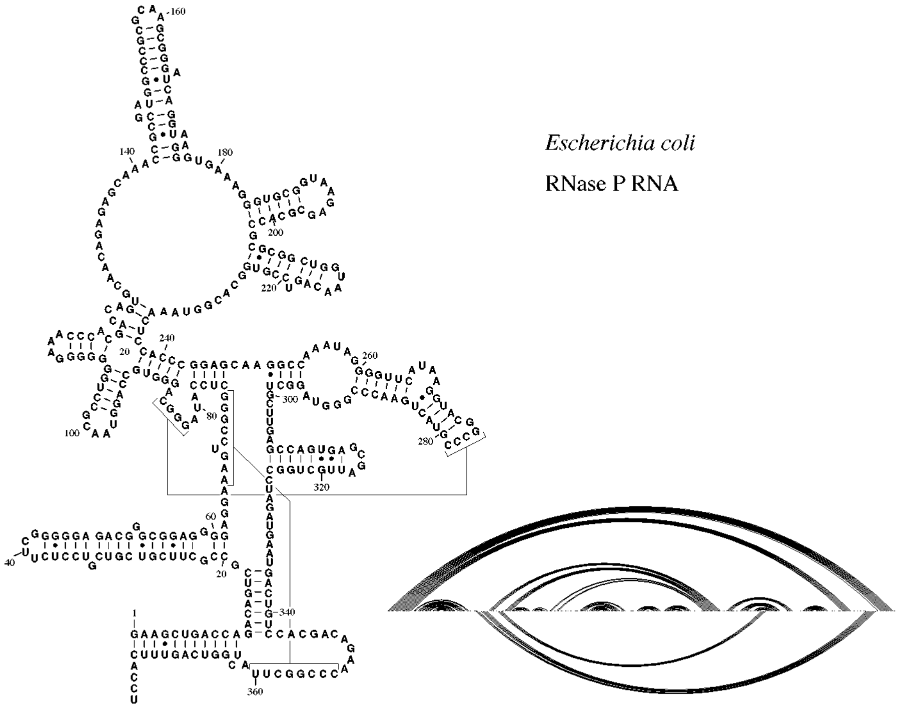

Figure 1. The contact structure of *Escherichia coli* RNAse P RNA contains two pseudo-knots [`http://jwbrown.mbio.ncsu.edu/RNaseP/home.html`]. The conventional secondary structure is drawn on the l.h.s., the (four) regions forming the pseudo-knots are marked by braces, interaction regions are connected. The arc diagram of the same structure is obtained by arranging the backbone along a line and indicating base pairs by arcs connecting the corresponding bases. The base pairs of the conventional secondary structure are drawn above the line, the two pseudo-knot stems are shown below the back-bone. For details see Section 2.

functional RNAs such as RNAseP RNA (Loria and Pan, 1996) and ribosomal RNA (Konings and Gutell, 1995) contain pseudo-knots. The diversity of molecular biological functions performed by pseudo-knots can be subdivided into three groups. Pseudo-knots at the 5′-end of mRNAs appear to adopt a role in the control of mRNA translation. For instance, the expression of replicase is controlled in several viruses either by ribosomal frame shifting (Ten Dam *et al.*, 1990; Brierley *et al.*, 1991; Dinman *et al.*, 1991; Chamorro *et al.*, 1992; Tzeng *et al.*, 1992) or by in-frame read-through of stop codons (Wills *et al.*, 1991). Both mechanisms involve pseudo-knots. Core pseudo-knots are necessary to form the reaction center of ribozymes. Most of the enzymatic RNAs with core pseudo-knots, such as RNAseP, are involved in cleavage or self-cleavage reactions (Michel and Westhof, 1990; Forster and Altman, 1990; Brown, 1991; Haas *et al.*, 1991). Pseudo-knots in the tRNA-like motifs at the 3′-end of the genomic RNA mediate replication control in several groups of plant viral RNA (Mans *et al.*, 1991).

It is important, therefore, to include pseudo-knotted structures into investigations of RNA sequence–structure relationships. In particular, we need to know whether the findings (i) through (vi) described above remain true when pseudo-knots are taken into account. Assertion (i), the existence of more sequences than structures, is a necessary prerequisite for all subsequent statements concerning the sequence-structure map of RNA. It is necessary therefore to estimate the number of RNA structures *with pseudo-knots* in order to decide whether the results quoted above can in fact be true for 'real' RNA molecules.

In the following two sections we give a detailed mathematical analysis of what we call bi-secondary structures. In a nutshell, bi-secondary structures generalize to the notion of secondary structures to include pseudo-knots without allowing overly involved knotted structures or nested pseudo-knots. In fact, almost all known pseudo-knotted structures, with the notable exception of the E. coli αmRNA, fall into this class.

In Section 2 we review a variety of equivalent graph-theoretical characterizations of bi-secondary structures and provide a way of efficiently determining whether a list of base pairs corresponds to a bi-secondary structure. Then we briefly review a few graph invariants that might be useful for determining the complexity of higher-order structures beyond the realm of bi-secondary structures. At the end of Section 2 we show that a convenient distance measure for comparing secondary structures can be used also in the presence of pseudo-knots (Section 2.7). In Subsection 2.8 we argue that the *intersection theorem* is valid for general nucleic acid contact structures. We say that an RNA sequence is *compatible* with a structure s if it can in principle form this structure irrespective of energetic constraints. This means that for each base pair (i, j) in s the sequence positions x_i and x_j are one of the six possible RNA base pairs **AU**, **UA**, **GC**, **CG**, **GU**, or **UG**. The set of sequences that actually fold into a given structure s is therefore a subset of the set of compatible sequences. The intersection theorem (Reidys *et al*., 1997) now states that for any two structures s and s' there are sequences which are compatible with both of them. This result is the reason why very different structures with very closely related sequences (Schuster *et al*., 1994) can exist. The fact that the intersection theorem holds for structures with pseudo-knots means that we have to expect shape space covering provided the fraction of neutral mutations is large enough (Reidys *et al*., 1997).

In Section 3 we determine the number of different structures with pseudo-knots. Combinatorial aspects of RNA secondary structures have been studied in detail by Waterman and co-workers (Stein and Waterman, 1978; Waterman, 1978; Waterman and Smith, 1978a, b; Penner and Waterman, 1993; Schmitt and Waterman, 1994; Waterman, 1995) and Hofacker *et al*. (1999). Using different techniques we give analytical upper bounds on the number of different bi-secondary structures showing that their number does not increase *much* faster than the number of secondary structures. The analytical results are complemented by numerical data (see Table 2 at the end of Section 3) indicating that the number S_n of 'reasonable' bi-secondary structures with chain length n grows approximately as $S_n \sim 2.35^n$. 'Reasonable' means here that the structures have no isolated base pairs (i.e., the minimum stack size is $l = 2$) and that hairpin loops contain at least $m = 3$ unpaired bases. For comparison, the number of secondary structures without pseudo-knots grows like 1.86^n. Exhaustive enumeration for short sequences suggest that only 1.65^n different secondary structures appear as minimum energy structures of sequences of length n (Grüner *et al*., 1996a). Hence the number 4^n of RNA sequences of length n, is much larger than the number of possible structures, independent of whether or not one takes pseudo-knots into account.

This observation poses the question *how* the sequences that fold into a given structure are distributed in sequence space. In Section 4 we describe a set of numerical experiments strongly suggesting that the inclusion of pseudo-knots does not alter the qualitative picture [properties (i) through (vi) above] of the RNA sequence–structure map. A short discussion (Section 5) concludes this contribution.

Readers who are not interested in the mathematical details of defining, characterizing, and counting contact structures of various types might want to skip Sections 2 and 3.

2. Contact Structures, Diagrams and Book-embeddings

2.1. ***Diagrams and diagram graphs.*** The three-dimensional structure of a linear biopolymer, such as RNA, DNA, or a protein can be approximated by its *contact structure*, i.e., by the list of all pairs of monomers that are spatial neighbors. Contact structures of polypeptides have been introduced by Ken Dill and co-workers in the context of lattice models of protein folding (Chan and Dill, 1988; Chen and Dill, 1995). They arise implicitly in knowledge-based potentials for polypeptides such as the Delauney–Tesselation potential described in Singh *et al*. (1996). Last but not least, RNA secondary structures form a special class of contact structures. The purpose of this section is to bring together different mathematical approaches that can be used to describe biopolymer structures: contact graphs, linked diagrams, book embeddings, and graph colorings.

A contact structure is represented by the *contact matrix* $\mathbf{C}$ with the entries $\mathbf{C}_{ij} = 1$ if the monomers i and j are spatial neighbors without being adjacent along the backbone, and $\mathbf{C}_{ij} = 0$ otherwise. Hence $\mathbf{C}_{ij} = 0$ if $|i - j| \leq 1$. We shall use the notation $[n] = \{1, \dots, n\}$.

We define a *diagram* $([n], \Omega)$ to consists of n vertices labeled 1 to n and a set Ω of *arcs* that connect non-consecutive vertices.

A closely related class of diagrams which also allow arcs between consecutive vertices are the *linked diagrams* introduced by Touchard (1952). These are studied in some detail in Hsieh (1973), Kleitman (1970), Stein (1978) and Stein and Everett (1978).

It is customary to arrange the vertices along the x-axis and to draw the vertices in such a way that they are confined in either the upper or the lower half-plane. The diagram of a contact structure with contact matrix **C** has the set of arcs

$$\Omega = \{\{i, j\} | \mathbf{C}_{ij} = 1\}. \tag{1}$$

The contact matrix is thus the adjacency matrix of the corresponding diagram.

With each diagram we may associate a *diagram graph* Γ with the following properties:

(i) The $n + 1$ vertices of Γ are labeled $0, 1, \dots, n$.
(ii) Γ contains the Hamiltonian cycle $[0, 1, \dots, n, 0]$.
(iii) The 'root' vertex 0 has degree 2.

Let **B** be the adjacency matrix of the backbone, i.e., the matrix with the entries $\mathbf{B}_{i,i+1} = \mathbf{B}_{i+1,i} = 1$, $i = 0, \dots, n-1$, and $\mathbf{B}_{0n} = \mathbf{B}_{n0} = 1$. Then the adjacency matrix of a diagram graph with $n + 1$ vertices is of the form

$$\mathbf{A} = \mathbf{B} + \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{C} \end{pmatrix}. \tag{2}$$

Equation (2) establishes the 1-1 correspondence of diagrams and the associated diagram graphs.

Essentially the same construction can be used for contact structures of molecules with a circular backbone, i.e., for circular ssRNA or ssDNA. The only restriction is that $\{1, n\}$ cannot be an arc in the case of a circular molecule. It is convenient in this case to define the corresponding diagram graph without the artificial root 0. Each graph Γ with a Hamiltonian cycle is then the diagram graph of a contact structure with a circular backbone. The results in the following discussion hold for both linear and circular nucleic acids.

DEFINITION 1. *A diagram is a* 1-diagram *if for any two arcs $\alpha, \beta \in \Omega$ holds $\alpha \cap \beta = \emptyset$ or $\alpha = \beta$.*

A diagram Δ is a 1-diagram if and only if associated diagram graph $\Gamma(\Delta)$ has vertex degrees less or equal to 3. Such graphs are often called *sub-cubic* or trivalent. The diagram graphs of 1-diagrams are closely related to cubic Hamiltonian graphs. The latter are studied in detail in Section 9.4 of Wagner and Bodendiek (1990): a graph S is *homeomorphic from* a graph Γ if S can be produced from Γ by inserting vertices of degree 2 into some edges of Γ. S is also called a *subdivision* of Γ. Obviously each cubic Hamiltonian graph gives rise to a diagram graph on n vertices by subdividing the edges of a Hamiltonian cycle. On the other hand, not all diagram graphs are homeomorphic from a cubic Hamiltonian graph: suppose $\{1, 3\}$ is an arc and 2 is an unpaired vertex. The corresponding diagram graph cannot be cubic because the triangle 1, 2, 3 cannot be obtained from a cubic graph.

2.2. *Secondary structures.* The classical definition of secondary structures (Waterman, 1978) requires that each base interacts with at most one other nucleotide. Thus nucleic acid secondary structures are special types of 1-diagrams. The second defining condition is that arcs do not cross. In terms of the contact matrix this means: if $\mathbf{C}_{ij} = \mathbf{C}_{kl} = 1$ and $i < k < j$ then $i < l < j$. With the following notation we will find a simpler formulation of condition 2:

Let $\alpha = \{i, j\}$ with $i < j$ be an arc of a diagram. We write $\bar{\alpha} = [i, j] \subset \mathbb{R}$ for the associated interval. Two arcs of a diagram are *consistent* if they can be drawn in the same half-plane without crossing each other. Equivalently, two arcs $\alpha, \beta \in \Omega$ of a diagram are consistent if either one of the following four conditions is satisfied:

(i) $\bar{\alpha} \cap \bar{\beta} = \emptyset$.
(ii) $\bar{\alpha} \subseteq \bar{\beta}$.
(ii) $\bar{\beta} \subseteq \bar{\alpha}$.
(iv) $\bar{\alpha} \cap \bar{\beta} = \{k\}$, a single vertex.

Case (iv) is ruled out by definition in 1-diagrams. The non-crossing condition thus may be expressed as follows: whenever the intervals of two arcs $\{i, j\}$ and $\{k, l\}$ have non-empty intersection then one is contained in the other (Schmitt and Waterman, 1994). Equivalently, we may simply define that *a secondary structure is a 1-diagram in which any two arcs are consistent.*

As a consequence, each secondary structure can be encoded as a string s of length n in the following way: if the vertex i is unpaired, then $s_i =$ ‘`.`’. Each arc $\alpha = \{p, q\}$ with $p < q$ translates to $s_p =$ ‘`(`’ and $s_q =$ ‘`)`’. As the arcs are consistent their corresponding parentheses are either nested, `(( ))`, or next to each other, `()()`. As there are no arcs between neighboring vertices in a 1-diagram there is at least one dot contained within each parenthesis. A variant of this notation is the *mountain representation* of RNA secondary structures (Hogeweg and Hesper, 1984). The ‘dot-parenthesis’ notation is used as a convenient notation in input and output of the `Vienna RNA Package`, a piece of public domain software for folding and comparing RNA molecules (Hofacker *et al*., 1994).

2.3. *Book-embedding of graphs.* A graph that can be embedded in the plane (or, equivalently on the sphere) is called *planar*. If it can be embedded in the plane in such a way that all its vertices lie on the exterior region it is called *outer-planar*. This class of graphs was introduced and characterized in terms of subgraphs in Chartrand and Harary (1967) and Sysło (1979). Clearly, a 1-diagram Δ is a secondary structure if and only if its diagram graph $\Gamma(\Delta)$ is outer-planar. The outer-planar embedding corresponds to the ‘circle representation’ of secondary structures.

A similar procedure leads to book-embeddings. A p-book is a set of p distinct half-planes (the *pages* of the book) that share a common boundary line ℓ, called the *spine* of the book. An embedding of a graph Γ into a book $\mathcal{B}$ consists of an ordering of the vertices along the spine of the book together with an assignment of each edge to a page of the book, in which edges assigned to the same page do not cross. The *book-thickness* (sometimes also called the page-number) $\mathrm{bt}(\Gamma)$ of a graph Γ is the minimal number p of pages of a book into which it can be embedded (Bernhart and Kainen, 1979). Book-embeddings have a practical application in the context of VLSI design. For an overview see Chung *et al*. (1987) and Heath *et al*. (1992).

Not surprisingly, the book thickness is closely related to other embedding properties of graphs. Below we list a few important results:

(i) $\mathrm{bt}(\Gamma) = 0$ if and only if Γ is a path.
(ii) $\mathrm{bt}(\Gamma) \leq 1$ if and only if Γ is outer-planar (Bernhart and Kainen, 1979).
(iii) $\mathrm{bt}(\Gamma) \leq 2$ if and only if Γ is a subgraph of a planar Hamiltonian graph (Bernhart and Kainen, 1979). Such graphs are sometimes called subhamiltonian.
(iv) $\mathrm{bt}(\Gamma) \leq 4$ if Γ is planar (Yannakakis, 1988).
(v) $\mathrm{bt}(K_n) = \lceil n/2 \rceil$, where K_n is the complete graph with n vertices (Bernhart and Kainen, 1979).
(vi) $\mathrm{bt}(K_{mn}) = \min(\lceil n/2 \rceil, \lceil m/2 \rceil)$, where K_{mn} is the complete bipartite graph with $m + n$ vertices.
(vii) $\mathrm{bt}(\Gamma) \leq \frac{3}{2}\sqrt{n} + 6$ for sub-cubic graphs (Chung *et al*., 1987).
(viii) $\mathrm{bf}(\Gamma) \leq O(\sqrt{m})$ if Γ is a graph with m edges(Malitz, 1994, b) .
(ix) $\mathrm{bf}(\Gamma) \leq O(\sqrt{g})$ if Γ is a graph of genus g (Malitz, 1994, b). (The genus of a graph is the minimum number of 'handles' one needs to add to a sphere so that the graph can be embedded on the resulting surface without crossing edges.)

The book thickness of a variety of other graph classes has been studied in detail, among them hypercubes (Chung *et al*., 1987), De Bruijn graphs (Obrenić, 1993), and various types of network graphs of practical interest (Games, 1986).

2.4. *The inconsistency graph of a diagram.*

DEFINITION 2. *Let* $\Delta = ([n], \Omega)$ *be a diagram. The* inconsistency graph $\Theta(\Delta)$ *of the diagram has vertex set* Ω *and* $\{\alpha, \beta\}$ *is an edge of* $\Theta(\Delta)$ *if and only if the arcs* α *and* β *are inconsistent in* Δ.

Essentially the same construction is used for the investigation of cubic Hamiltonian graphs in Wagner and Bodendiek (1990). We shall see that the inconsistency graph is a useful construction for characterizing embedding properties of diagram graphs.

THEOREM 1. *Let* Δ *be a diagram. Then the following statements are equivalent.*

(i) The diagram Δ *can be drawn without intersecting arcs.*
(ii) The diagram graph $\Gamma(\Delta)$ *is planar.*
(iii) The inconsistency graph $\Theta(\Delta)$ *is bipartite.*
(iv) $\Gamma(\Delta)$ *has a 2-page book embedding.*

Proof. (i $\Longleftrightarrow$ ii) Δ can be drawn without intersection arcs if and only if $\Gamma(\Delta)$ is planar because the Hamiltonian cycle $\mathcal{H}$ of $\Gamma(\Delta)$ divides the plane into the interior and the exterior of $\mathcal{H}$ which correspond to the upper and lower half-plane of the diagram Δ, respectively.

(ii $\Longleftrightarrow$ iii) can be shown in the same way as the analogous result for cubic Hamiltonian graphs in Wagner and Bodendiek (1990), see also Even and Itai (1971). (ii $\Longleftrightarrow$ iv) follows immediate from Bernhart and Kainen (1979, Theorem 2.5) as a planar diagram graph is by construction Hamiltonian.

As noted in Even and Itai (1971), the determination of the book thickness of a Γ is equivalent to finding a minimal vertex-coloring of a certain circle graph, which in our case is the intersection graph $\Theta(\Delta)$. This problem is in general NP-complete. The following observation simplifies the task by reducing the number of arcs that have to be considered.

Two arcs $\alpha = \{i, j\}$ and β are *stacked* if $\beta = \{i-1, j+1\}$ or $\beta = \{i+1, j-1\}$. A *stem* is a subset Ψ of arcs α_0 through α_h such that α_p and α_{p+1} are stacked for $p = 0, \dots, h-1$. It is easy to show that the arcs of a stem Ψ of a 1-diagram are either all isolated vertices or they are contained in the same component of the inconsistency graph $\Theta(\Delta)$. Furthermore, all arcs of a stem have the same adjacent vertices in $\Theta(\Delta)$. We may therefore use a reduced intersection graph $\hat{\Theta}(\Delta)$, the vertices of which are the stems. (In addition, we may recursively remove vertices of degree 2 that are not contained in a triangle before forming the intersection graph. This has the effect of removing bulges and interior loops that interrupt stems.) Examples of reduced intersection graphs are given in Figs 3 and 4.

Most of the literature on linked diagrams deals with *complete* diagrams, that is, each vertex $x \in [n]$ is incident with an arc (Touchard, 1952; Kleitman, 1970; Stein, 1978). It is straightforward to extend Touchard's definition of reducible diagrams to the incomplete diagrams considered here:

DEFINITION 3. *A diagram* $([n], \Omega)$ *is* reducible *if there exists an interval* $[p, q] \subset [n]$ *such that*

(i) *For each* $\alpha \in \Omega$ *holds either* $\alpha \cap [p, q] = \emptyset$ *or* $\alpha \subseteq [p, q]$.
(ii) *There is an arc* $\alpha \in \Omega$ *such that* $\alpha \cap [p, q] = \emptyset$.
(iii) *There is an arc* $\alpha \in \Omega$ *such that* $\alpha \subseteq [p, q]$. *If a diagram is not reducible, it is* irreducible.

The following equivalence is proved in Haslinger (1997):

LEMMA 1. *A diagram* Δ *is irreducible if and only if its inconsistency graph* $\Theta(\Delta)$ *is connected. A sub-diagram corresponds to one or more components of the inconsistency graph.*

Reducible diagrams can therefore be viewed as being composed of substructures. These substructures do not in general conform to the conventional decomposition into stems and loops of an RNA that forms the basis of the standard energy model of nucleic acid secondary structures (Freier *et al*., 1986).

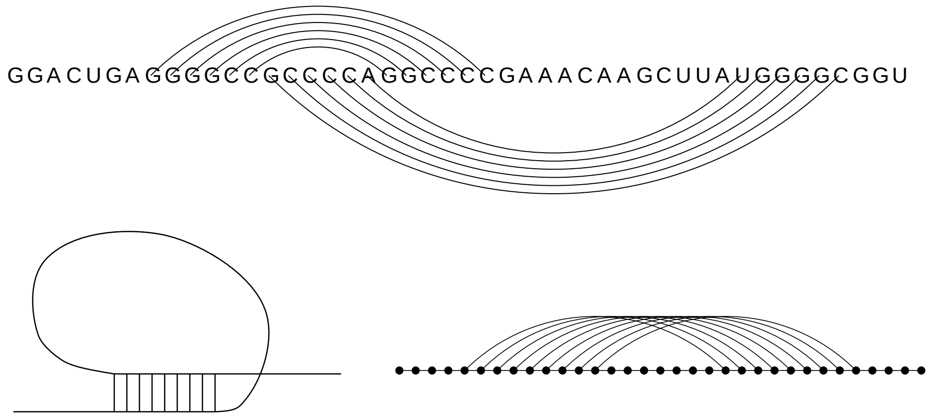

Figure 2. The contact structure of the proposed SRV-1 frameshift signal contains a pseudo-knot, see Ref. Ten Dam *et al*. (1994). Pseudo-knots such as this one belong to the class of bi-secondary structures. Knots such as the one in the lower part of the figure do not belong to the class of bi-secondary structures. Knots, in contrast to pseudo-knots, may contain parallel stranded helices which so far have not been described for RNA.

2.5. *Bi-secondary structures.*

DEFINITION 4. *A* bi-secondary structure *is a 1-diagram that can be drawn in the plane without intersections of arcs.*

We may draw the arcs in the upper or lower half-plane, but they are not allowed to intersect the x-axis. In other words, it can be embedded in 2-page book. Bi-secondary structures are therefore ‘superpositions’ of two secondary structures.

The virtue of bi-secondary structures is that they capture a wide variety of RNA pseudo-knots, [Figs 1 and 2 (upper part)], while at the same time they exclude true knots. Knotted RNAs could in principle arise either from parallel stranded helices (Fig. 2), or in very large molecules from sufficiently complicated cross-linking patterns. Parallel-stranded RNA has not been observed (so far), see, however, Fortsch *et al*. (1996) on parallel-stranded DNA. Wollenzien Cantor *et al*. (1980) have searched unsuccessfully for knots in large RNAs. The definition of bi-secondary structures, by allowing a planar drawing of the structure, rules out both possibilities.

Among the RNA structures with pseudo-knots, almost all are bi-secondary structures. Our examples include several viral RNAs such as Coronavirus (Brierley *et al*., 1991), Luteovirus (Ten Dam *et al*., 1990), and Retrovirus RNA (Chamorro *et al*., 1992), as well as catalytic RNAs such as RNAseP RNA (Loria and Pan, 1996), tmRNA (Vlassov *et al*., 1995; Felden *et al*., 1997), and ribosomal RNAs (Gutell *et al*., 1994). We have encountered only a single exception, namely αmRNA (Tang and Draper, 1990).

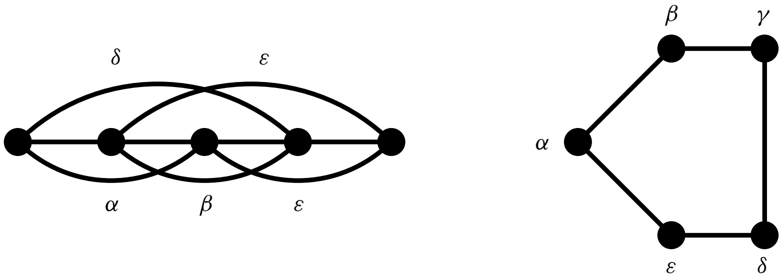

Figure 3. Theorem 2 is not valid for general diagrams. The inconsistency graph of the diagram Δ_5 is a pentagon and hence is neither bipartite nor does it contain a triangle.

THEOREM 2. *Let Δ be a 1-diagram. Then the following statements are equivalent:*

(i) Δ *is a bi-secondary structure.*
(ii) $\Gamma(\Delta)$ *is planar.*
(iii) $\Theta(\Delta)$ *is bipartite.*
(iv) $\Gamma(\Delta)$ *has a 2-page book-embedding.*
(v) Among any three arcs of Δ *at least two are consistent.*
(vi) $\Theta(\Delta)$ *does not contain a triangle.*

Proof. The equivalence of (i), through (iv) is proved in Theorem 1 for all diagrams. The equivalence of (v) and (vi) follows immediately from the definition of $\Theta(\Delta)$. The implication (iii$\Longrightarrow$v) is obvious. Finally, it is possible to show that $\neg$(ii) implies $\neg$(v) based on Kuratowski's (1930) theorem. For the details we refer to Haslinger (1997).

The practical importance of Theorem 2 lies in the fact that existence or non-existence of triangles in $\Theta(\Delta)$ can be checked very easily, and hence we have a very efficient (polynomial time) method for deciding whether a diagram Δ is a bi-secondary structure or not. Note that the equivalence of (iii) and (vi) does not hold for general diagrams. A counterexample is shown in Fig. 3.

Being the union of the two secondary structures $([n], \Omega_U)$ and $([n], \Omega_L)$ we can represent each bi-secondary structure as a string s using two types of parentheses: as in a secondary structure we write a dot '`.`' for all unpaired vertices. A pair $\{p, q\} \in \Omega_U$ becomes $s_p =$ '`(`' and $s_q =$ '`)`', while an arc $\{p, q\} \in \Omega_L$ becomes $s_p =$ '`[`' and $s_q =$ '`]`'. Unfortunately, the decomposition of a bi-secondary structure into two secondary structures is in general not unique, see Fig. 4.

The fact that $\Theta(\Delta)$ is bipartite allows us to define a *normal form* for this representation by means of the following rule: the leftmost arc of each connected component of $\Theta(\Delta)$ belongs to Ω_U. In particular, all isolated vertices of $\Theta(\Delta)$ are contained in Ω_U. The normal form of a secondary structure therefore contains only dots and (round) parentheses. Within each non-trivial connected component of $\Theta(\Delta)$ the distribution of arcs between Ω_U and Ω_L is unique because the component

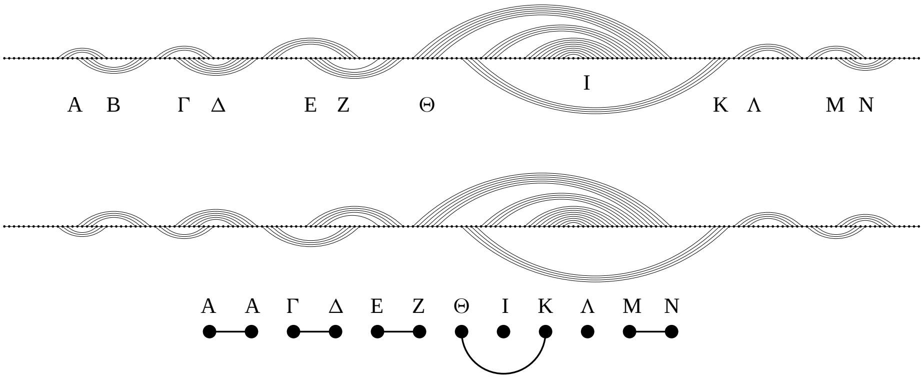

Figure 4. Two diagrams encoding the 3′ non-coding region of tobacco mosaic virus RNA (Abrahams *et al*., 1990). The upper diagram corresponds to the normal form, the lower diagram maximizes the number of upper arcs. Stems are labeled by uppercase greek letters. The third line shows the inconsistency graph (see Section 2.4) of the tmvRNA structure. It is bipartite and hence, by Theorem 2, the tmvRNA structure is a bi-secondary structure.

is bipartite. All arcs in a stack have a common neighboring vertex in $\Theta(\Delta)$, hence they all belong to the same class of the partition. Therefore, in normal form, all arcs belonging to the same stack are written with the same type of brackets.

2.6. *Beyond bi-secondary structures.* The following example shows that there are natural RNA structures that are more complicated than bi-secondary structures. The *Escherichia coli* α-operon mRNA folds into a structure that is required for allosteric control of translational initiation (Tang and Draper, 1990). Compensatory mutations have defined an unusual pseudo-knotted structure (Tang and Draper, 1989), the thermodynamics of which were subsequently investigated in detail (Gluick and Draper, 1994). The diagram of its contact structure cannot be drawn without intersections, see Fig. 5. To our knowledge it is the only known RNA structure that cannot be embedded in a 2-page book.

In this subsection we briefly discuss a few graph properties that could be used for a classification of polymer structure complexity beyond the realm of bi-secondary structures. Clearly, one may use its book thickness. A closely related quantity is the chromatic number of the intersection graph: a *color partition* of a graph Γ is partition $V = V_1 \cup V_2 \cup \cdots \cup V_c$ of its vertex set into c subsets V_i such that no two vertices in V_i are adjacent. The *chromatic number* $\chi(\Gamma)$ is the smallest number c of colors for which a color partition of Γ can be found.

An arbitrary diagram Δ can be decomposed into substructures by means of the following obvious result: let $\Delta = ([n], \Omega)$ be a diagram and let $\mathcal{V} : \Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_c$ be a partition of the set of arcs. Then the sub-diagram $([n], \Omega_i)$, $i = 1, \ldots, c$, can be drawn without intersection if and only if $\mathcal{V}$ is a color partition of

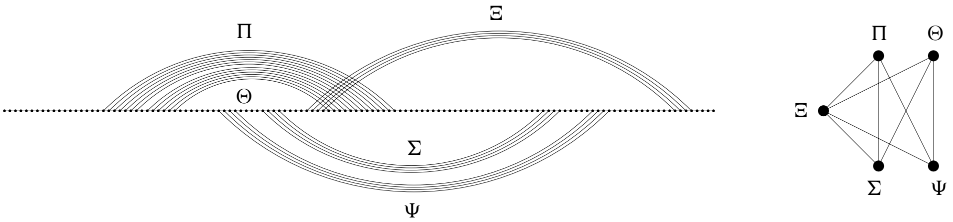

Figure 5. Diagram of the contact structure of $E.\ coli$ α-mRNA. The structure contains five stems, labeled by uppercase Greek letters. We may choose the color partition if $\Theta(\Delta)$ such that all arcs in a stem have the same color. It therefore suffices to draw the inconsistency graph for stems (r.h.s. of the figure). It contains triangles, thus the diagram of this RNA structure is not a bi-secondary structure. It is easy to check that $\chi(\Theta(\Delta)) = 3$.

the inconsistency graph $\Theta(\Delta)$. Noticing that $\chi(\Gamma) = 1$ if Γ contains no edges and $\chi(\Gamma) = 2$ if Γ is bipartite with non-empty edge set, the following characterization follows immediately:

(i) Δ is a secondary structure iff $\chi(\Theta(\Delta)) = 1$;
(ii) Δ is a bi-secondary structure iff $\chi(\Theta(\Delta)) \le 2$.

Clearly, $\chi(\Theta(\Delta))$ equals the minimum number of pages of all book embeddings in which the the ordering of the vertices along the spine coincided with the natural ordering along the backbone. In general, we have $\mathrm{bt}(\Gamma(\Delta)) \le \chi(\Theta(\Delta))$ for all diagrams. We remark that graphs with moderate chromatic numbers can be characterized by results similar to Kuratowski's theorem for planar graphs. For instance, one can show for $k \le 4$, that a graph Γ with chromatic number $\chi(\Gamma) \ge k$ contains a subdivision of the complete graph K_k (Dirac, 1952). The generalization of this proposition to $k > 4$ is known as Hajós' conjecture. It is false for $k \ge 7$ and unsolved for $k = 5$ and $k = 6$ (Holton and Sheehan, 1993). It seems that $\chi(\Theta(\Delta))$ is in fact the more useful quantity, as there are no efficient algorithms to determine the book-thickness of a given graph, and $\chi(\Theta(\Delta))$ accounts for the immutable ordering of the backbone vertices, whereas the book-thickness might decrease by changing this ordering.

A quite different algebraic graph invariant μ, introduced by de Verdière (1990), leads to the same classification of structures for small μ:

$\mu = 1$ $\Gamma(\Delta)$ is a circle, Δ has no arcs.
$\mu \le 2$ $\Gamma(\Delta)$ is outer-planar, Δ is a secondary structure.
$\mu \le 3$ $\Gamma(\Delta)$ is planar, Δ is a bi-secondary structure.

The graphs with $\mu \le 4$ have recently been identified as the *flat* or *linklessly embeddable* graphs (Lovász and Schrijver, 1996). A useful characterization of this class of graphs is proved in Robertson *et al*. (1995, b): 'A graph is non-flat if and only if it has no minor in the so-called Petersen family'. The graph V_8^*, Fig. 6, is a valid diagram graph. It is easy to check that V_8^* is flat and that its inconsistency graph is

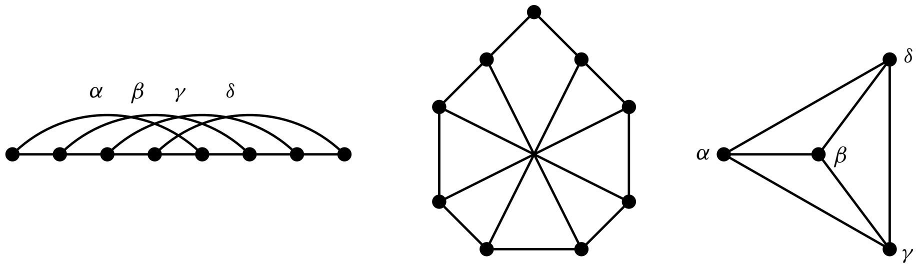

Figure 6. The graph V_8^* and its inconsistency graph.

$\Theta(V_8^*) = K_4$. Hence there are flat diagram graphs for which $\chi(\Theta(\Delta)) \geq 4$. Thus there is no direct correspondence between $\chi(\Theta(\Delta))$ and μ, not even for 1-diagrams.

2.7. ***A metric for 1-diagrams.*** An interesting algebraic interpretation of secondary structures was proposed in Reidys and Stadler (1996). Interpreting each arc $\{i, j\}$ as a transposition (i, j) on $[n]$ we may assign the permutation

$$\pi(\Delta) = \prod_{\alpha\in\Omega}(i_\alpha, j_\alpha) \tag{3}$$

to each diagram Δ. One observes: (i) if Δ a 1-diagram then $\pi(\Delta)$ is an involution. (ii) An involution π is the permutation representation of a 1-diagram if and only if its cycle decomposition does not contain a canonical transposition, i.e., a transposition of the form $(i, i+1)$. (iii) Different 1-diagrams give rise to different involutions.

A natural set of generators for the symmetric group S_n is the set $\mathcal{T}$ of all transpositions. The corresponding length function is

$$\ell(\pi) = n - \mathrm{cyc}(\pi), \qquad \pi \in S_n, \tag{4}$$

where $\mathrm{cyc}(\pi)$ is the number of cycles into which π decomposes. We have $\ell(\tau) = 1$ if and only if $\tau \in \mathcal{T}$ is a transposition. The associated metric is the canonical metric on the Cayley graph $\Gamma(S_n, \mathcal{T})$, see Reidys and Stadler (1996) for a detailed discussion. As the involutions form a subset of S_n we have

THEOREM 3. *The function*

$$d(\Delta, \Delta') = \ell(\pi(\Delta)\pi(\Delta')^{-1}) = n - \mathrm{cyc}(\pi(\Delta)\pi(\Delta')^{-1}), \tag{5}$$

where $\pi(\Delta)$ denotes the permutation representation of a diagram Δ, is a metric on the set of all 1-diagrams with n vertices.

In particular, two 1-diagrams Δ and Δ' have distance $d(\Delta, \Delta') = 1$ if and only if they differ by a single arc. Metrics on 'shape space' are necessary for a detailed quantitative study of sequence–structure maps. Applications to RNA secondary structures are reported for instance in Fontana *et al.* (1993a) and Schuster *et al.* (1994).

2.8. *The Intersection Theorem.* The virtue of equation (3) is not limited to defining a metric on the set of structures. Suppose we are given an alphabet of monomers (for instance {**A**, **U**, **G**, **C**} in the case of RNA) and a rule that determines which pairs of monomers may form a base pair (**AU**, **UA**, **GC**, **CG**, **GU**, **UG** in the case of RNA).

DEFINITION 5. *A sequence s is* compatible *with a structure (1-diagram) Δ if for each arc $\{i, j\}$ the letters (monomers) s_i and s_j fulfill the pairing rule. The set of all sequences that are compatible with Δ is denoted by* $\mathbf{C}[\Delta]$.

THEOREM 4. *(Intersection Theorem) Let Δ and Δ' be 1-diagrams. Then $\mathbf{C}[\Delta] \cap \mathbf{C}[\Delta']$ is non-empty.*

The proof of this result in Reidys *et al*. (1997) is valid for all 1-diagrams, not only for secondary structures. The intersection theorem sets the stage for shape space covering: it allows close-by sequences to fold into structures that are as different as desired — given a suitable folding potential. Further applications of equation (3) can be found in Weber (1997).

3. COMBINATORICS

3.1. *Enumeration 1-diagrams.* The number X_n of all diagrams on n vertices is $X_n = 2^{(n-1)(n-2)/2}$ as there are $(n-1)(n-2)/2$ possible arcs (Söler and Jankowski, 1991), which can be arbitrarily combined to form a diagram.

In Section 2.7 we have shown that all 1-diagrams correspond to involutions, therefore the number T_n of involutions on $[n]$ is an upper bound for the number D_n of 1-diagrams on $[n]$. The combinatorics of involutions is discussed for instance in the book by Wilf (1994):

PROPOSITION 1. *The number T_n of involutions fulfills the recursion*

$$T_n = T_{n-1} + (n-1)T_{n-2} \qquad n \geq 2 \qquad \textit{and} \qquad T_0 = T_1 = 1\,,$$

and has the asymptotic form

$$T_n \sim \frac{1}{\sqrt{2}} n^{n/2} \exp\left(-\frac{n}{2} + \sqrt{n} - \frac{1}{4}\right) .$$

The number of involutions T_n therefore grows faster than exponential in the sense that $\sqrt[n]{T_n} \rightarrow \infty$. 1-Diagrams can be counted by a very similar recursion as the following result shows:

THEOREM 5. *The number of 1-diagrams fulfills the recursion*

$$D_{n+2} = D_{n+1} + (n+1)D_n - D_{n-1} + D_{n-2} \quad n \geq 2$$

$$D_0 = D_1 = D_2 = 1, \;\; D_3 = 2.$$

Proof. The first few values of D_n are obvious, $D_0 = 1$ is a convenient definition. The recursion is derived as follows: a 1-diagram on $n+2$ vertices can be formed either by adding a lone vertex to a 1-diagram on $n+1$ vertices or by adding an arc $\{1, k\}$ to a 1-diagram Δ on n vertices by inserting the vertex labeled k between the $k-1$st and the kth vertex of Δ. Note, however, that Δ must be a 1-diagram, but in addition it might have an arc $\{k-1, k\}$ in Δ, as these vertices are separated by the endpoint of the newly introduced arc in the new structure. Viewing this differently, we may either add the arc $\{1, k\}$ or the Ψ-like structure consisting of the arcs $\{1, k\}$ and $\{k-1, k+1\}$, which leaves us with a 1-diagram on $n-2$ vertices and the same problem. Repeating this argument we arrive at the following expansion:

Hence we have $D_{n+2} = D_{n+1} + nD_n + (n-1)D_{n-2} + (n-3)D_{n-4} + \cdots$. Observing that D_{n+1} can of course be written in the same form and substituting into the above equations yields

$$D_{n+2} = (n+1)D_n + nD_{n-1} + (n-1)D_{n-2} + (n-2)D_{n-3} + \cdots + 2D_1 + D_0 - D_{n-1}\,.$$

Subtracting the corresponding expansion for D_{n+1} yields

$$D_{n+2} - D_{n+1} = (n+1)D_n - D_{n-1} + D_{n-2}\,.$$

A simple rearrangement now completes the proof.

COROLLARY 1. $\lim_{n\to\infty} \sqrt[n]{D_n} = \infty$.

Proof. The series D_n is obviously monotonically increasing. Hence the series $a_{n+2} = (n+1)a_n, a_0 = a_1 = 1$ is a lower bound. It is well known that $a_n = (n-1)!!$ grows faster than exponentially.

REMARK 1. A very similar formula is obtained for the case of a circular backbone. There are D_{n-2} diagrams with arc $\{1, n\}$ on n vertices. Thus the number of 1-diagrams with circular backbone is $D'_n = D_n - D_{n-2}$.

An exponential upper bound can be found, however, on the numbers $D_n(c)$ of 1-diagrams whose inconsistency graph has chromatic $\chi(\Theta(\Delta)) \le c$. We find

THEOREM 6. $D_n(c) \le (2c+1)^n$.

Proof. Consider a 1-diagram $\Delta = ([n], \Omega)$ with $\chi(\Theta(\Delta)) \le c$. Then there is a color partition of Ω with c colors. As $([n], \Omega_i)$ is a secondary structure, it can be encoded in dot-parenthesis notation. Coloring the parenthesis with a different color for each class Ω_i of the color partition hence yields a unique representation of Δ. This representation can be interpreted as a string of length n over an alphabet consisting of ‘ . ’ and c different pairs of brackets, i.e., with $2c+1$ letters.

Theorem 6 is not a very good estimate as we shall see in Section 3.3.

Table 1. The constants A_{ml} in equation (7) for secondary structures without pseudo-knots.

m	l		
	1	2	3
1	2.618	1.986	1.716
2	2.414	1.899	1.680
3	2.289	1.849	1.652
5	2.147	1.783	1.612

3.2. ***Secondary structures.*** A secondary structure on $n+1$ digits may be obtained from a structure on n digits either by adding a free end at the right-hand end or by inserting a base pair $1 \equiv (k+2)$. In the second case the substructure enclosed by this pair is an arbitrary structure on k digits, and the remaining part of length $n-k-1$ is also an arbitrary valid secondary structure. Therefore, we obtain the following recursion formula for the number S_n of secondary structures:

$$\begin{aligned} S_{n+1} &= S_n + \textstyle\sum_{k=m}^{n-1} S_k S_{n-k-1}, \qquad n \geq m+1 \\ S_0 &= S_1 = \cdots = S_{m+1} = 1. \end{aligned} \tag{6}$$

This expression has first been derived by Waterman (1978); m denotes the minimum number of unpaired digits in a hairpin loop. Similar recursions can be derived for the numbers $\Psi_n^{(m,l)}$ of secondary structures with minimum hairpin length m and minimum stack length l, see Hofacker *et al*. (1999) for details. Asymptotically, these numbers behave as

$$\Psi_n^{(m,l)} \sim B_{m,l} n^{-3/2} A_{m,l}^n \, . \tag{7}$$

The most important numbers are collected in Table 1. A more detailed table can be found in Hofacker *et al*. (1999).

Detailed combinatorial studies on various aspects of secondary structure graphs are based on equation (6), see for instance Penner and Waterman (1993), Stein and Waterman (1978), Waterman (1978, 1995), Waterman and Smith (1978a, b) and Hofacker *et al*. (1999). In the following we shall make use of the number

$$s(n,k) = \frac{1}{k}\binom{n-k}{k+1}\binom{n-k-1}{k-1} \tag{8}$$

of secondary structures of length n with k base pairs. This closed formula was recently derived in Schmitt and Waterman (1994).

3.3. ***Bi-secondary structure.*** A first naive upper bound is $D_n(2) \leq S_n^2$, because on each side of the x-axis we have a secondary structure on n vertices. Theorem 5 implies $D_n(2) \leq 5^n$. A slightly better bound can be derived using the enumeration of secondary structures:

Lemma 2. $D_n(2) \le \max\limits_{\substack{0\le k+l\le n/2\\ l\le k}} \frac{n}{2}\binom{n-k-1}{k-1}\binom{n-k}{k+1}\binom{n-2k}{2l}\binom{2l}{l}.$

Proof. We start with the $s(n,k)$ secondary structures with k arcs. In order to produce a bi-secondary structure we use $2l$ of the $n-2k$ unpaired positions for introducing l additional arcs. There are $\binom{n-2k}{2l}$ possible choices for these additional pairs, which may form any of the $C_l = \frac{1}{l+1}\binom{2l}{l}$ possible configurations of l matched parentheses. C_l is a Catalan number. Without losing generality we may assume that $l \le k$, i.e., the partial secondary structure with the larger number of pairs is drawn above the x-axis. Thus

$$D_n(2) \le \sum_{k=0}^{n/2}\sum_{l=0}^{k} s(n,k)\binom{n-2k}{2l} C_l\,.$$

Replacing the sums by appropriate multiples of the maximum entry is trivial.

Note that this bound is still a gross overestimate: (i) it contains all the redundancy of the `().[]`-representation. (ii) The number C_l also counts conformations of square brackets of the form `[]`, which do not correspond to a graph at all, and it counts conformations in which not all square brackets are inconsistent with an arc that is represented by a round bracket. These latter configurations are counted more than once.

Corollary 2. $\lim\limits_{n\to\infty}\sqrt[n]{D_n(2)} \le 4.76136931.$

Proof. Let $A_n(k,l)$ denote argument of the maximum in Lemma 2. It is straightforward to compute

$$\begin{aligned} A(x,y) &= \lim_{n\to\infty}\frac{1}{n}\log A_n(nx,xy)\\ &= 2(1-x)\log(1-x) - 2x\log x - (1-2x)\log(1-2x)\\ &\quad -(1-2x-2y)\log(1-2x-2y) - 2y\log(y). \end{aligned}$$

Set $A = \max\{\,A(x,y)\,|\,0 \le x+y \le 1/2 \wedge y \le x\,\}$. Then $\lim\sqrt[n]{D_n(2)} \le \exp(A)$. Solving the optimization problem that defines A is straightforward. A short computation shows that $\hat{y} = 1/\sqrt{21}$ and $\hat{x} = (7-\sqrt{21})/14$ is the only local maximum with $x, y \le 1/2$. It violates the condition $y \le x$, however. The solution thus lies on the boundary of the triangle $(0,0)$, $(1/2,0)$ and $(1/4,1/4)$. Setting $y = 0$ one obtains the maximum $\hat{x} = 1/2 - 1/\sqrt{20}$. Along the edge $x + y = 1/2$ we find $\hat{y} = 1/\sqrt{12}$ violating the condition $y \le x$. With $x = y$ we arrive at the cubic equation $31x^3 - 31x^2 + 10x - 1 = 0$ which has a single real solution $\hat{x} \approx 0.1942$. We find $A(\hat{x},\hat{x}) \approx 1.5605329 = A$, because this value is much larger than the values of $A(x,y)$ at the three corners of the triangle.

Table 2. Best estimates for the constant $A^{(2)}_{ml}$. The counting data were fitted by the model $a\, n^{-b}\, c^n$.

m	l		
	1	2	3
1	4.42	2.49	2.00
2	4.03	2.43	1.94
3	3.81	2.35	1.89
5	3.44	2.22	1.74

More sophisticated models of RNA take into account that (i) base pairs must enclose at least $m = 3$ other bases, and (ii) that isolated base pairs are energetically disfavored. In Hofacker *et al*. (1999) the numbers $\Psi_n^{(m,l)}$ of secondary structures with stack size at least l base pairs and separation of the vertices incident with an arc at least m is derived. We define $\Psi_n^{(m,l;\kappa)}$ to be the number of 1-diagrams with $\chi(\Theta(\Delta)) \le \kappa$ and with the same restrictions, and set

$$A^{(\kappa)}_{ml} = \lim_{n\to\infty} \sqrt[n]{\Psi_n^{(m,l;\kappa)}}. \tag{9}$$

Clearly we have $\Psi_n^{(m,l;2)} \le [\Psi_n^{(m,l)}]^\kappa$ because the 1-diagram Δ is a superposition of at most κ secondary structures. In particular, we find the upper bound $A^{(2)}_{3,2} \le 3.418$ for the biophysical case.

We have not been able to derive an exact counting series for bi-secondary structures. Hence we resorted to a numerical survey. We pursued three different strategies for estimating the number of bi-secondary structures:

(1) Complete enumeration is feasible only for very small values of n because the number of structures grows faster than 2^n.

(2) As an alternative we produce random strings from the alphabet `().[]` and check each string if it is the normal form of a bi-secondary structure. The number of secondary structures is then estimated by $5^n \times N_{\mathrm{nf}}/N_{\mathrm{sample}}$, where N_{sample} is the size of the random sample and N_{nf} is the number of detected normal forms in the sample.

(3) Using the recursion for secondary structures with given minimal stack length l and given minimal hairpin size m, described in detail in Tacker *et al*. (1996), we randomly generate a sample of pairs of secondary structures. Interpreting these as the upper and lower part of bi-secondary structure we check their superpositions for being normal forms of bi-secondary structures. The number of bi-secondary structures is then approximately $\Psi_n^{(m,l)} \times N_{\mathrm{nf}}/N_{\mathrm{sample}}$, where the numbers $\Psi_n^{(m,l)}$ of secondary structures with hairpins of length at least m and minimal stack length m can be obtained recursively, see Hofacker *et al*. (1999).

Our best estimates are compiled in Table 2. In the biologically interesting case, $m = 3$ and $l = 2$, we find $A^{(2)}_{3,2} \approx 2.35$. Judging from the exhaustive enumeration data (Grüner *et al.*, 1996a) we should expect that the number of structures that actually occur as minimum energy structures is still smaller.

4. Statistics

4.1. ***A simplified energy model for pseudo-knots.*** In order to incorporate pseudo-knots into secondary structure computations we first have to devise an energy model. Naturally, we require that this energy function extends the standard model for RNA secondary structures without pseudo-knots.

The standard energy model is based on decomposing a secondary structure into its 'loops' (Zuker and Sankoff, 1984). For secondary structures without pseudo-knots this decomposition is unique and coincides with the so-called minimum cycle basis of the secondary structure graph (Leydold and Stadler, 1998). The free energy of a particular secondary structure is computed as the sum of the contributions of the individual loops. These contributions depend on the type of the loop (stacked base pairs, hairpin loop, bulge, interior loop, or multi-branch loop), its size, and on the sequence of nucleotides, see e.g., Walter *et al.* (1994).

We emphasize that the energy model for pseudo-knotted structures introduced in this section is not intended as an accurate potential for predicting pseudo-knots in particular (biologically relevant) sequences. It is intended as a simplified model that allows us to investigate the likelihood of pseudo-knots in an ensemble of sequences and the stability of pseudo-knots against point mutations of the sequence. It is shown in Tacker *et al.* (1996) for (pseudo-knot-free) RNA secondary structures that such statistical properties are surprisingly robust against changes in the parameter set and the choice of the folding algorithm. For instance, most global properties of RNA folding are already present in the 'maximum matching' model, which, instead of an elaborate energy model, simply seeks to maximize the number of base pairs (Tacker *et al.*, 1996). A potential function that captures the most salient features of pseudo-knots is therefore sufficient for our purposes.

Very little experimental information is available on the thermodynamics of pseudo-knots, see, however, Wyatt *et al.* (1990). On the other hand, the geometric constraints of RNA structures are well understood (Saenger, 1984; Pleij *et al.*, 1985). Hence we start from the following three principles:

(i) Loops that are not involved in pseudo-knots have the same energy contributions as in pseudo-knot-free RNA secondary structures.
(ii) The stacking energies of base pairs are not affected by pseudo-knot formation even in stems that are part of pseudo-knots.
(iii) Steric hindrance is the major contribution to the pseudo-knot energies.

The energy parameters detailed in Walter *et al.* (1994), and implemented in release 1.2 of the `Vienna RNA Package` (Hofacker *et al.*, 1994), are used in this study for

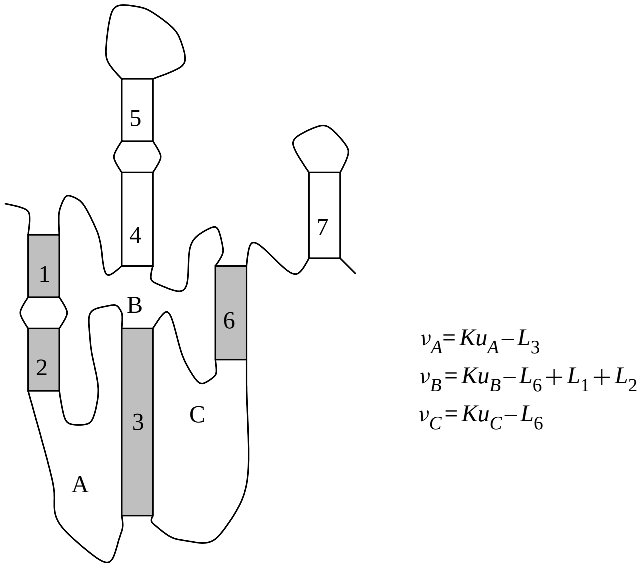

Figure 7. Schematic drawing of an RNA structure with pseudo-knots. The three loops A, B, C and the four stems 1, 2, 3, and 6 are involved in pseudo-knots. The evaluation of loops A and C is straightforward as they contain only a single paired region, namely stack 3. Three stems are contained in loop B; we assume that stack 6 is the longest one. The ν-parameters of the three pseudo-knotted loops are listed on the r.h.s. The energy contributions of base pair stacking and the contributions of all unmarked loops are evaluated according to the standard model.

the non-pseudo-knot contributions. The basic idea for parameterizing the pseudo-knot contributions rests on two simplifications: (i) RNA stacks are viewed as stiff rods and (ii) unpaired regions are assumed to be very flexible. Within a loop that is involved in pseudo-knot formation, we assume that each of the stacks formed by the pseudo-knotted base pairs is a stiff helix. This reasoning leads to an ansatz based upon the following quantities:

u = number of unpaired bases in the loop.
$L_{\max}$ = number of base pairs in the longest pseudo-knot stack.
L_i = number of bases in pseudo-knot stack i.
K = number of stacked base pairs that can be bridged by one unpaired base.

First we define a measure for the sterical hindrance in the pseudo-knotted loop:

$$\nu = Ku - L_{\max} + \sum_{i \neq \max}^{n} L_i \,. \tag{10}$$

This expression assumes that all other parts of a loop can be used to meet the constraint introduced by the longest stacked region $L_{\max}$ within the loop, see Fig. 7.

The free energy contributions of the unpaired regions can be estimated from a theory by Jacobson and Stockmeyer (1950). The same approach is used for long loops in the standard energy model for RNA secondary structures. If the free energy needed to join the ends of an unrestricted, zero volume polymer is known, the theory predicts the free energy needed to form a similar but larger loop. The minimum length of an RNA loop that behaves according to the Jacobson–Stockmayer theory is not known. We therefore introduce a parameter $\bar{\nu}$ and define the energy function as follows:

$$E(L) = \begin{cases} \infty & \text{if } \nu < 0 \\ E_{ps} & \text{if } 0 \leq \nu \leq \bar{\nu} \\ E_{ps} + \alpha \log(\nu/\bar{\nu}) & \text{if } \nu > \bar{\nu}. \end{cases} \tag{11}$$

Our energy model therefore has four free parameters that need to be estimated from the available experimental data, namely K, $\bar{\nu}$, E_{ps} and α. For simplicity we fixed α at the same value that is used for all non-pseudo-knotted loops: $\alpha = 1078.56$ cal mol^{-1} (at 37°C).

4.2. ***Folding.*** Given the sequence, one can compute the secondary structure with the minimum energy by means of dynamic programming (Waterman, 1978; Zuker and Sankoff, 1984). In the presence of pseudo-knots this is no longer true. In the present study we use Tacker's kinetic folding algorithm (Tacker *et al*., 1996) which is based on (Martinez, 1984). It first produces a list of all possible stems of a given sequence and then determines the free energies of the loops and stacks. The most stable stem is the first one added to the folding structure. Using this as a constraint, we compile a list of the remaining possible stems and add the most stable one to the growing structure. This procedure is repeated until the free energy of the structure cannot be decreased anymore.

The parameters K, $\bar{\nu}$, and E_{ps} are adjusted by predicting the structures of a sample of sequences that are known to form pseudo-knots. This set includes seven fragments with about 80 nt from bacteriophages that form H-type pseudo-knots, *E. coli* tmRNA containing five pseudo-knots, and RNAse P sequences from several different species [for details see Haslinger (1997)]. The best results were obtained using $K = 4$, $\bar{\nu} = 9$, $E_{ps} = 4.2$ kcal mol^{-1}. The same value of E_{ps} was used in Abrahams *et al*. (1990). In order to check the influence of these parameters on the sequence–structure relation of RNA we also used a parameter set leading to an unrealistically large number of predicted pseudo-knots in the test sequences ($K = 3$, $\bar{\nu} = 10$, $E_{ps} = 2.0$ kcal mol^{-1}).

4.3. ***The sequence–structure map with pseudo-knots.*** The average number of base pairs and related statistical properties of the predicted structures depend very little on the inclusion of pseudo-knots and the choice of the pseudo-knot parameters. This is not surprising as the relative stability of base pairs and unpaired regions remains essentially unchanged. The average loop size decreases with the 'unrealistic' pseudo-knot potential because loop regions may take part in pseudo-knots at very little entropic cost.

Table 3. Average Number of Pseudo-knots per Structure.

Potential	30	50	70	100	Slope
Realistic	0.009	0.020	0.028	0.038	0.000408
Exaggerated	0.150	0.311	0.450	0.628	0.006784

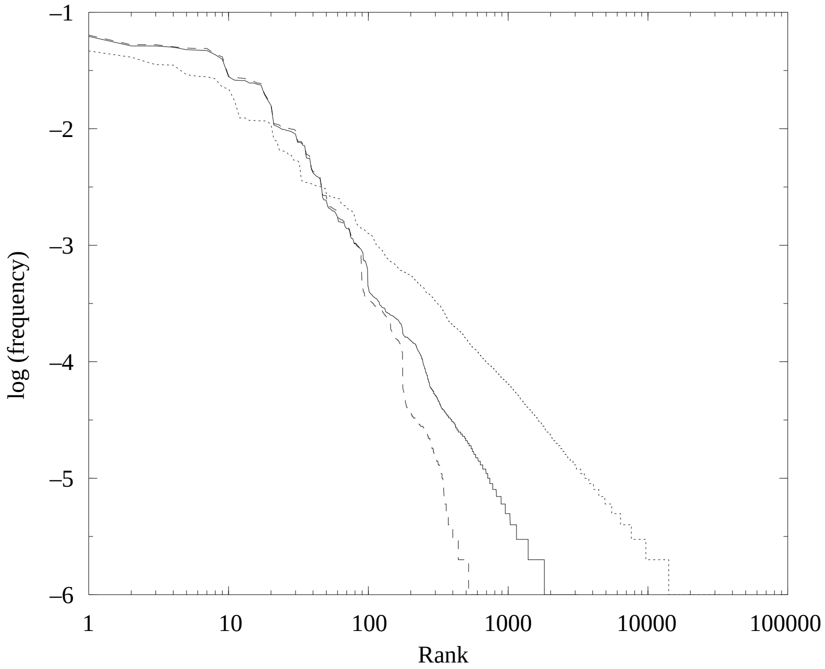

Figure 8. The frequency distribution of RNA contact structures. Shapes are ranked by their frequencies. The particular example shown here deals with the loop structures (Shapiro and Zhang, 1990) of 1 million RNA molecules of chain length $n = 100$ which are derived from the contact structures by further eliminating all details concerning stack lengths and loop sizes. The full line is the distribution for the realistic pseudo-knot potential, the dashed line refers to secondary structures only, and the dotted line to the exaggerated pseudo-knot potential. While the inclusion of pseudo-knots somewhat increases the fraction and the diversity of rare structures it does not change the general shape of the distribution.

The frequency of pseudo-knots in random sequences is tabulated in Table 3. For the realistic potential we find a pseudo-knot every ∼2500 bases, while with the exaggerated potential one would expect one pseudo-knot in every random sequence of length $n = 148$.

As we have seen in the previous section, there are still many more sequences than structures. In order to obtain a better impression of the relationship between the numbers of sequences and structures that arise through folding, we determine the *rank order statistics* of folded structures. To this end we compute the structures of a large number of randomly chosen sequences and rank them according to their frequency f of occurrence in the sample. A plot of log f versus the logarithm of

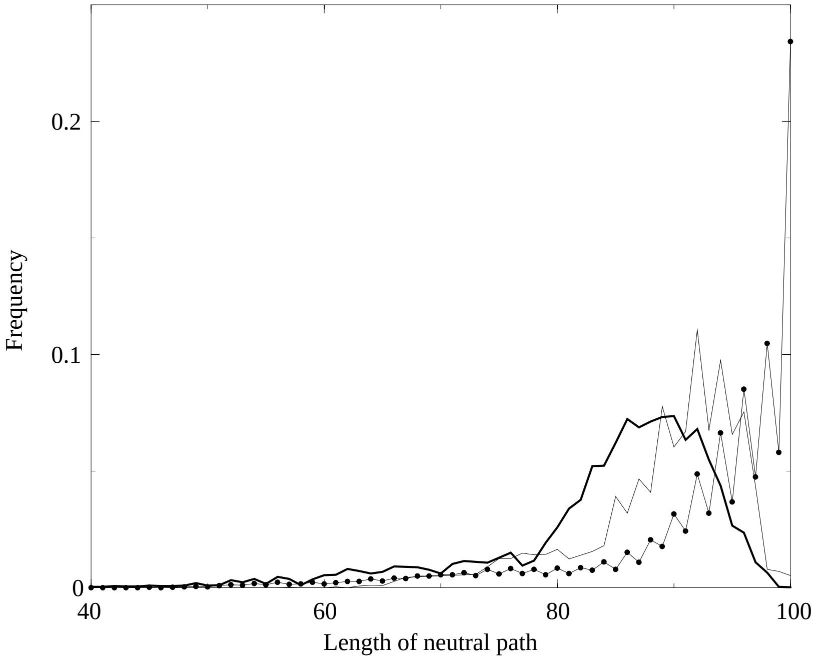

Figure 9. Lengths of neutral paths were determined for a large sample of initial sequences with $n = 100$. The average lengths are $\mathcal{L} = 92.25$ in the absence of pseudo-knots (black dots), $\mathcal{L} = 89.69$ for the realistic pseudo-knot potential (thin line), and $\mathcal{L} = 87.22$ for the exaggerated version (thick line).

the rank reveals a generalized Zipf's law (Zipf, 1949), Fig. 8. While the inclusion of pseudo-knots somewhat increases the fraction and the diversity of rare structures (large ranks) it does not change the general shape of the distribution. As for 'pure' secondary structures there is only a small number of common structures into which almost all sequences fold.

Naturally, we ask how sequences folding into the same (common) secondary structure are distributed in sequence space. We call the set $S(\psi)$ of all sequences (genotypes) folding into phenotype (contact structure) ψ the *neutral set* of ψ. More precisely, $S(\psi)$ is the pre-image of ψ w.r.t. the folding map algorithm. As for 'pure' secondary structures, a large fraction λ of point mutations is neutral, i.e., does not change the structure. On the other hand, RNA sequences folding into a particular structure are not significantly clustered: they form a percolating network spanning the entire sequence.

The fraction λ of neutral point mutations was estimated from 6000 independently generated random sequences, see Table 4. As observed in Grüner *et al*. (1996a, b), we find that λ decreases somewhat with chain length (the large values for $n = 30$ being caused in part by the large number of short sequences that 'fold' into the open structure). The fraction of neutral neighbors approaches an asymptotic value slightly above 0.5. Surprisingly, this value is almost independent of the potential

Table 4. Fraction λ of neutral mutants.

Potential	$n = 30$	$n = 50$	$n = 70$	$n = 100$	$n = 200$	∞
Secondary structures	0.708	0.628	0.604	0.578	0.548	0.523
Realistic	0.709	0.625	0.598	0.576	0.550	0.529
Exaggerated	0.680	0.588	0.547	0.540	0.533	0.529

function: even a potential leading to a large fraction of pseudo-knotted structures decreases λ only by a few percent.

A random graph theory (Reidys *et al*., 1997; Reidys, 1997) shows that there is threshold value of about $\lambda^* = 0.307$ (for a 4-letter alphabet). If the fraction of neutral neighbors exceeds this threshold, then the set of all sequences folding into a given structure s forms a single connected network, which has been termed the *neutral network* of s.

These neutral networks can be conveniently detected by means of a simple computer experiment. A *neutral path* starts at a randomly chosen sequence. Then we construct a series of subsequent mutants such that each sequence along the path folds into the same structure as the initial sequence, and such that each step increases the Hamming distance from the starting point. The strict logic on base pairing in RNA makes it necessary to consider two types of mutations: (i) point mutations in the unpaired regions of the molecules, and (ii) the substitution of one possible base pair (**GC**, **CG**, **GU**, **UG**, **AU**, **UA**). All other mutations in paired regions necessarily change the structure, for instance by changing a **GU** pair into a **GG** mismatch. If there are neutral networks in sequence space the neutral path will reach a length $\mathcal{L}$ close to n before there is no neutral mutant further away from the starting point (n is the maximal Hamming distance between sequences of length n). On the other hand, if the neutral sets $S(\psi)$ form isolated clusters we will find $\mathcal{L} \ll n$. When interpreting the lengths of neutral paths we have to keep in mind that (i) the search procedure only produced lower bounds on the diameter of neutral networks, and (ii) that a pair of random sequences has an expected distance of $0.75n$ for a 4-letter alphabet. The data in Fig. 9 are therefore a clear indication for the existence of percolating neutral networks in the presence of pseudo-knots.

5. Discussion

Secondary structures form a particular class of contact structures. In this contribution we have considered a natural generalization of this class. Indeed, most known RNA structures with pseudo-knots are bi-secondary structures (which do not involve nested pseudo-knots). Bi-secondary structures correspond to planar graphs while secondary structures form the sub-class of outer-planar graphs.

The inconsistency graph introduced in Section 2.4 is a useful construction capturing most of the geometrical features of nucleic acid structure. Its chromatic number may serve as a measure of structural complexity. It seems possible that an analogous construction will be useful for classifying and comparing protein structures as well. The analysis of graph-theoretical properties of classes of contact structures might also be useful for designing energy models that are more realistic and/or algorithmically easier than pair potentials. The standard folding potential for RNA and DNA secondary structures, for instance, is based on loops, that is, induced subgraphs of the diagram graph that are circles. The total energy of a secondary structure is defined as the sum of the sequence-dependent energy contributions of all loops [see, e.g., Freier *et al*. (1986)]. It is by no means obvious how this energy function should be generalized to include non-secondary structure features such as pseudo-knots, **G**-quartets, or knots, because in general there is no unique decomposition of a graph into loops.

In order to understand the sequence–structure mapping of a class of biopolymers it is necessary to have bounds on the number of structures that can possibly be formed for a given set of sequences. We can expect the existence of neutral networks and shape space covering only if the number of sequences by far exceeds the number of structures. While the number of possible contact structures grows faster than exponentially with the length of the molecules we find exponential upper bounds when the structural complexity is limited. In particular, there are not more than some 4.7^n possible bi-secondary structures. If we enforce in addition the sterical (loop-length at least 3) and thermodynamic (no isolated base pairs) constraints of natural RNA sequences, then this bound drops to 3.42^n. Exhaustive enumeration indicates that the actual number of bi-secondary structures with biophysical constraints grows roughly as 2.35^n. Therefore the number of RNA sequences, 4^n, exceeds by far the number of possible bi-secondary structures.

We have then devised a simple energy function extending the standard model to incorporate pseudo-knots. Our ansatz assumes that steric hindrance is the major contribution to pseudo-knot energies counteracting the stabilizing effect of the additional base pairings. Based on this approach we used a kinetic folding procedure to show that the inclusion of pseudo-knots does not significantly change the global features of the sequence structure map of RNA: there are many more sequences than structures, and almost all sequences fold into one of a small number of common structures. Common structures are uniformly distributed over sequence space.

Neutral networks in sequence space can therefore be modeled as random graphs (Reidys *et al*., 1997). This ansatz generalizes from secondary structures to 1-diagrams without modifications. The only input parameter in this model, namely the fraction λ of neutral neighbors, has been determined computationally. Computer simulations agree with the prediction of a random graph theory: the fraction of neutral mutations, $\lambda > 0.5$, is well above the threshold value of $\lambda^* \approx 0.306$, hence all sequences folding into a given common structure form a single percolating network that spans the entire sequence space. This is verified by the detection of neutral paths that extend through the entire sequence space.

The intersection theorem is valid for bi-secondary structures, hence the random graph approach (Reidys *et al.*, 1997), can be used to predict the relative locations of the neutral networks of two different common structures. In particular, we have to expect shape space covering, i.e., the neutral networks of any two common structure come very close to each other at least in some parts of the sequence space. This sets the stage for the evolutionary transitions between different structures described in detail in Weber (1997) and Fontana and Schuster (1998).

In summary, the mathematical results and the computer simulations presented in this contribution indicate that pseudo-knots do not change the qualitative picture of the RNA sequence–structure map as it was obtained from studying secondary structures.

ACKNOWLEDGEMENTS

Stimulating discussions with Ivo L. Hofacker and Christoph Flamm are gratefully acknowledged.

REFERENCES

Abrahams, J. P., M. van den Berg, E. van Batenburg and C. Pleij (1990). Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.* **18**, 3035–3044.

Bernhart, F. and P. C. Kainen (1979). The book thickness of a graph. *J. Comb. Theor.* **B27**, 320–331.

Bonhoeffer, S., J. S . McCaskill, P. F. Stadler and P. Schuster (1993). RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur. Biophys. J.* **22**, 13–24.

Brierley, I., N. J. Rolley, A. J. Jenner and S. C. Inglis (1991). Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol* **229**, 889–902.

Brown, J. W. (1991). Structure and evolution of ribonuclease P RNA. *Biochemie* **73**, 689–697.

Cantor, C. R., P. L. Wollenzien and J. E. Hearst (1980). Structure and topology of 16S ribosomal RNA. an analysis of the pattern of psoralen crosslinking. *Nucl. Acids Res.* **8**, 1855–1872.

Chamorro, M., N. Parkin and H. E. Varmus (1992). An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proc. Natl. Acad. Sci. USA* **89**, 713–717.

Chan, H. S. and K. A Dill (1988). Interchain loops in polymers: effects of excluded volume. *J. Chem. Phys.* **90**, 492–508.

Chartrand, G. and F. Harary (1967). Planar permutation graphs. *Ann. Inst. Henri Poincaré* **B3**, 433–438.

Chen, S.-J. and Ken A. Dill (1995). Statistical thermodynamics of double-stranded polymer molecules. *J. Chem. Phys.* **103**, 5802–5808.

Chung, F. R. K., F. T. Leighton and A. L. Rosenberg (1987). Embedding graphs in books: A layout problem with applications to VLSI design. *SIAM J. Alg. Disc. Math.* **8**, 1987.

Dinman, J. D., T. Icho and R. B. Wickner (1991). A-1 ribosomal frameshifting in a double-stranded RNA virus of yeast forms a gag-pol fusion protein. *Proc. Natl. Acad. Sci. USA* **88**, 174–178.

Dirac, G. A. (1952). A property of 4-chromatic graphs and some remarks on critical graphs. *J. London Math. Soc.* **27**, 85–92.

Even, S. and A. Itai. (1971). Queues, stacks, and graphs, in *Theory of Machines and Computation*, Z. Kohavi and A. Paz, (Eds), New York: Academic Press, pp. 71–86.

Felden, B., H. Himeno, A. Muto, J. P. McCutcheon, J. Atkins and R. F. Gesteland (1997). Probing the structure of the Escherichia coli 10Sa RNA (tmRNA). *RNA* **3**, 89–103.

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

Fontana, W., D. A. M. Konings, P. F. Stadler and P. Schuster (1993a). Statistics of RNA secondary structures. *Biopolymers* **33**, 1389–1404.

Fontana W. and P. Schuster (1998). Continuity in evolution: on the nature of transitions. *Science* **280**, 1451–1455.

Fontana, W., P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger and P. Schuster (1993b). RNA folding and combinatory landscapes. *Phys. Rev.* **E47**, 2083–2099.

Forster, A. C. and S. Altman (1990). Similar cage-shaped structures for the RNA component of all ribonuclease P and ribonuclease MRP enzymes. *Cell* **62**, 407–409.

Fortsch, I., H. Fritzsche, E. Birch-Hirschfeld, E. Evertsz, R. Klement, T. M. Jovin and C. Zimmer (1996). Parallel-stranded duplex DNA containing dA·dU base pairs. *Biopolymers* **38**, 209–220.

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

Games, R. (1986). Optimal book embeddings of FFT, benes, and barrel shifter networks. *Algorithmica* **1**, 233–250.

Gluick, T. C. and D. E. Draper (1994). Thermodynamics of folding a pseudoknotted mRNA fragment. *J. Mol. Biol.* **241**, 246–262.

Grüner, W., R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler and P. Schuster (1996a). Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Monath. Chem.* **127**, 355–374.

Grüner, W., R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler and P. Schuster (1996b). Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering. *Monath. Chem.* **127**, 375–389.

Gutell, R. R., N. Larsen and C. R. Woese (1994). Lessons from an evolving rRNA: 16S and 23S rRNA from a comparative perspective. *Microbiol. Rev.* **58**, 10–26.

Haas, E. S., D. P. Morse, J. W. Brown, J. F. Schmidt and N. R. Pace (1991). Long-range structure in ribonuclease P RNA. *Science* **254**, 853–856.

Haslinger, Christian (1997). RNA secondary structures with pseudoknots. Master's thesis, Inst. f. Theoretische Chemie, University of Vienna, 1997.

`http://www.tbi.univie.ac.at/ papers/Masters_theses.html`.

Heath, L. S., F. T. Leighton and A. L. Rosenberg (1992). Comparing queues and stacks as mechanisms for laying out graphs. *SIAM J. Discr. Math.* **5**, 398–412.

Hofacker, I. L., W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacke, and P. Schuster (1994). Fast folding and comparison of RNA secondary structures. *Monath. Chem.* **125**, 167–188.

Hofacker, I. L., P. Schuster and P. F. Stadler (1999). Combinatorics of RNA secondary structures. *Discr. Appl. Math.* **89**, 177–207.

Hogeweg, P. and B. Hesper (1984). Energy directed folding of RNA sequences. *Nucl. Acid. Res.* **12**, 67–74.

Holton, D. A. and J. Sheehan (1993). *The Petersen Graph*, in *Australian Mathematical Society Lecture Series* **7**, Cambridge: Cambridge University Press.

Hsieh, W. N. (1973). Proportions of irreducible diagrams. *Studies in Appl. Math.* **52**, 277–283.

Huynen, M. A. (1996). Exploring phenotype space through neutral evolution. *J. Mol. Evol.* **43**, 165–169.

Huynen, M. A., P. F. Stadler and W. Fontana (1996). Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA* **93**, 397–401.

Jacobson, H. and W. H. Stockmeyer (1950). Intramolecular reaction in polycondensations. *J. Chem. Phys.* **18**, 1600–1606.

Kleitman, D. (1970). Proportions of irreducible diagrams. *Studies in Appl. Math.* **49**, 297–299.

Konings, D. A. M. and R. R. Gutell (1995). A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA* **1**, 559–574.

Kuratowski, K. (1930). Sur le problème des courbes gauches en topologie, *Fund. Math.* **15**, 271-283.

Leydold, J. and P. F. Stadler (1998). Minimal cycle bases of outerplanar graphs. *Elec. J. Comb.* **5**, R16. See `http://www.combinatorics.org`.

Loria, Andrew and T. Pan (1996). Domain structure of the ribozyme from eubacterial ribonuclease P. *RNA* **2**, 551–563.

Lovász, L. and A. Schrijver (1996). The Colin de Verdière number of linklessly embeddable graphs. preprint.

Malitz, S. M. (1994a). Genus g graphs have pagenumber $o(\sqrt{g})$. *J. Algorithms* **17**, 85–109.

Malitz, S. M. (1994b). Graphs with e edges have pagenumber $o(\sqrt{E})$. *J. Algorithms* **17**, 71–84.

Mans, R., C. W. A. Pleij and L. Bosch. (1991). Transfer RNA-like structures: structure, function and evolutionary significance. *Eur. J. Biochem.* **201**, 303–324.

Martinez, H. M. (1984). An RNA folding rule. *Nucl. Acid. Res.* **12**, 323–335.

Michel, F. and E. Westhof (1990). Modelling of the three-dimmensional architecture of group I catalytic introns based on comparative sequence anaysis. *J. Mol. Biol.* **216**, 585–610.

Obrenić, Bojana. (1993). Embedding De Bruijn graphs and shuffle-exchange graphs in five pages. *SIAM J. Discr. Math.* **6**, 642–654.

Penner, R. C. and M. S. Waterman (1993). Spaces of RNA secondary structures. *Adv. Math.* **101**, 31–49.

Pleij, C. W., K. Rietveld and L. Bosch (1985). A new principle of RNA folding based on pseudoknotting. *Nucl. Acid. Res.* **13**, 1717–1731.

Reidys, C. M. (1997). Random induced subgraphs of generalized n-cubes. *Adv. Appl. Math.* **19**, 360–377.

Reidys, C. and P. F. Stadler (1996). Bio-molecular shapes and algebraic structures. *Comp. & Chem.* **20**, 85–94.

Reidys, C., P. F. Stadler and P. Schuster (1996). Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.* **59**, 339–397.

Robertson, N., P. Seymore and R. Thomas (1995a). Petersen family minors. *J. Comb. Theory* **B64**, 155–184.

Robertson, N., P. Seymore and R. Thomas (1995b). Sachs' linkless embedding conjecture. *J. Comb. Theory* **B64**, 185–227.

Saenger, W. (1984). *Principles of Nucleic Acid Structure*, London: Springer Verlag.

Schmitt, W. R. and M. S. Waterman (1994). Linear trees and RNA secondary structure. *Discr. Appl. Math.* **12**, 412–427.

Schuster, P. (1995). How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnol.* **41**, 239–257.

Schuster, P., W. Fontana, P. F. Stadler and I. L. Hofacker (1994). From sequences to shapes and back: A case study in RNA secondary structures. *Proc. R. Soc.* **B255**, 279–284.

Shapiro, B. A. and K. Zhang (1990). Comparing multiple RNA secondary structures using tree comparisons. *CABIOS* **6**, 309–318.

Singh, R. K., A. Tropsha and I. I. Vaisman (1996). Delaunay tessellation of proteins: Four body nearest neighbor propensities of amino acid residues. *J. Comput. Biol.* **3**, 213–221.

Söler, F. and K. Jankowski (1991). Modeling RNA secondary structures I. Mathematical structural model of predicting RNA secondary structures. *Math. Biosc.* **105**, 167–190.

Stein, P. R. (1978). On a class of linked diagrams, I. Enumeration. *J. Comb. Theory* **A24**, 357–366.

Stein, P. R. and C. J. Everett (1978). On a class of linked diagrams. II. Asymptotics. *Disc. Math.* **22**, 309–318.

Stein, P. R. and M. S. Waterman (1978). On some new sequences generalizing the Catalan and Motzkin numbers. *Disc. Math.* **26**, 261–272.

Sysło, M. M. (1979). Characterizations of outerplanar graphs. *Discr. Math.* **26**, 47–53.

Tacker, Manfred, W. Fontana, P. F. Stadler and P. Schuster (1994). Statistics of RNA melting kinetics. *Eur. Biophys. J.* **23**, 29–38.

Tacker, M., P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker and P. Schuster. Algorithm independent properties of RNA structure prediction. *Eur. Biophy. J.* **25**, 115–130.

Tang, C. K. and D. E. Draper (1989). An unusual mRNA pseudoknot structure is recognized by a protein translation repressor. *Cell* **57**, 531–536.

Tang, C. K. and D. E. Draper (1990). Evidence for allosteric coupling between the ribosome and repressor binding sites of a translationally regulated mRNA. *Biochemistry* **29**, 4434–4439.

Ten Dam, E., I. Brierly, S. Inglis and C. Pleij (1994). Identification and analysis od the pseudoknot-containing *gag-pro* ribosomal frameshift signal of simian retrovirus-1. *Nucl. Acids Res.* **22**, 2304–2310.

Ten Dam, E. B., C. W. A. Pleij and L. Bosch (1990). RNA pseudoknots and translational frameshifting on retroviral, coronaviral and luteoviral RNAs. *Virus Genes* **4**, 121–136.

Touchard, J. (1952). Sur une problème de configurations et sur les fractions continues. *Canad. J. Math.* **4**, 2–25.

Tzeng, T. H., C. L. Tu and J. A. Bruenn (1992). Ribosomal frameshifting requires a pseudoknot in the saccharomyces cerevisiae double-stranded RNA virus. *J. Virology* **66**, 999–1006.

de Verdière, Y. C. (1990). Sur un novel invariant des graphes et un critère de planarité. *J.*

Comb. Theory **B50**, 11–21.

Vlassov, V. V., G. Zuber, B. Felden, J. P. Behr and R. Griege (1995). Cleavage of tRNA with imidazole and spermine imidazole constructs: a new approach for probing RNA structures. *Nucl. Acid. Res.* **23**, 3161–3167.

Wagner, K. and R. Bodendiek (1990). *Graphentheorie II*, Mannheim: B.I. Verlag.

Walter, A. E., D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews and M. Zuker (1994). Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of RNA folding. *Proc. Natl. Acad. Sci. USA* **91**, 9218–9222.

Waterman, M. S. (1978). Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Stud.* **1**, 167 – 212.

Waterman, M. S. (1995). *Introduction to Computational Biology: Maps, Sequences, and Genomes*, London: Chapman and Hall.

Waterman, M. S. and T. F. Smith (1978a). Combinatorics of RNA hairpins and cloverleaves. *Studies Appl. Math.* **60**, 91–96.

Waterman, M. S. and T. F. Smith (1978b). RNA secondary structure: a complete mathematical analysis. *Math. Biosc.* **42**, 257–266.

Weber, J. (1997). Dynamics on Neutral Evolution. PhD thesis, Friedrich Schiller University, Jena, January 1997. `http://www.tbi.univie.ac.at/ papers/PhD_theses.html`.

Westhof, E. and L. Jaeger (1992). RNA pseudoknots. *Current Opinion Struct. Biol.* **2**, 327–333.

Wilf, H. S. (1994). *Generating functionology*, San Diego, CA: Academic Press.

Wills, N., R. F. Gesteland and J. F. Atkins (1991). Evidence that a downstream pseudoknot is required for translational readthrough of the moloney murine leukemia virus gag stop codon. *Proc. Natl. Acad. Sci. USA* **88**, 6991–6995.

Wyatt, J. R., J. D. Puglisi and I. T. Tinoco (1990). RNA pseudoknots stability and loop size requirements. *J. Mol. Biol* **214**, 455–470.

Yannakakis, M. (1988). Embedding planar graphs in four pages. *J. Comput. Syst. Sci.* **38**, 36–67.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*, Reading, MA: Addison-Wesley.

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

Received 3 February 1998 and accepted 29 October 1998