RNA Structures with Pseudo-Knots

Graph-Theoretical and Combinatorial Properties

Peter F. Stadler[a,b,*] and Christian Haslinger[a]

[a]Institut für Theoretische Chemie, Universität Wien, Austria

[b]The Santa Fe Institute, Santa Fe, New Mexico, U.S.A.

*Mailing Address: Peter F. Stadler
Inst. für Theoretische Chemie, Universität Wien
Währingerstr. 17, A-1090 Wien, Austria
Phone: **43 1 40480 665 Fax: **43 1 40480 660
E-Mail: studla@tbi.univie.ac.at or stadler@santafe.edu

Abstract

Secondary structures of nucleic acids are a particularly interesting class of contact structures. Many important RNA molecules contain pseudoknots, which are excluded explicitly by the definition of secondary structures. We propose here a generalization of secondary structures that incorporates "non-nested" pseudoknots. We also introduce a measure for the complexity of more general contact structures in terms of the chromatic number of their intersection graph. We show that RNA structures without nested pseudoknots form a special class of planar graphs. Upper bounds on their number are derived, showing that there are fewer different structures than sequences.

1. Introduction

Presumably the most important problem and the greatest challenge in present day theoretical biophysics is deciphering the code that transforms sequences of 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 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 "classical" definition of secondary structures excludes pseudoknots [47] mostly for technical reasons: The folding problem for RNA can be solved efficiently by dynamics programming [47, 50] in the absence of pseudoknots. The sequence structure relation of RNA was studied in detail in a series of papers [2, 6, 8, 7, 11, 12, 32, 37, 38] 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.
(iii) Sequences that fold into a "common" structure are distributed nearly uniformly in sequence space.
(iv) A sequence folding into "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 [12, 26].

(vi) Almost all "common" structure 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 [19, 31]: 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 [18]. Fast diffusion together with perpetual innovation makes these landscapes ideal for evolutionary adaptation.

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. Notably, functional RNAs such as RNAseP RNA [23] and ribosomal RNA [21] contain pseudoknots. It is an important question, therefore, whether the findings (i) through (vi) remain true when pseudoknots 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 pseudoknots* in order to decide whether the results quoted above can in fact be true for "real" RNA molecules.

Combinatorial aspects of RNA secondary structures have been studied in detail by Waterman and coworkers [25, 30, 36, 44, 46, 47, 45] and Hofacker *et al.* [14]. In section 2 we discuss generalizations of secondary structures and their graph-theoretical properties. The combinatorics of these objects is then studied by both analytical and numerical methods in section 3. A short discussion (section 4) concludes this contribution.

2. Contact Structures and Diagrams

2.1. Secondary Structures

The three-dimensional structure of a linear biopolymer, such as RNA, DNA, or a protein can be approximated by their *contact structure*, i.e., by the list of all pairs of monomers that are spatial neighbors. 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 $|i-j| \le 1$. We shall use the notation $[n] \overset{\text{def}}{=} \{1, \dots, n\}$.

Definition. A *diagram* $([n], \Omega)$ 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 allow also arcs between consecutive vertices are the *linked diagrams* introduced by Touchard [42]. These are studied in some detail in the references [17, 20, 34, 35].

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 $\mathbf{C}$ has the set of arcs

$$\Omega \overset{\text{def}}{=} \left\{ \{i,j\} \,\middle|\, \mathbf{C}_{ij} = 1 \right\}. \tag{1}$$

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

Definition. A *diagram graph* is a simple vertex labeled 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.

Lemma 1. There is an isomorphism $\imath$ between the diagrams with n vertices and the diagram graphs with $n+1$ vertices.

Proof. Let $\mathbf{B}$ be the matrix with the entries $\mathbf{B}_{i,i+1} = \mathbf{B}_{i+1,i} = 1$, $i = 0, \dots, n-1$, and $\mathbf{B}_{0n} = \mathbf{B}_{n0} = 1$. Furthermore let $\mathbf{C}$ be a symmetric $n \times n$ matrix with entries 0 or 1 and $\mathbf{C}_{ij} = 0$ whenever $|i-j| \le 1$. In other words $\mathbf{C}$ is the contact matrix of a diagram on n vertices. The adjacency matrix of a diagram graph with $n+1$ vertices is of the form

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

since the “root” vertex 0 has only the neighbors 1 and n. Since $\mathbf{B}$ is already defined by n, this decomposition is unique and $\mathbf{A}$ can be constructed if $\mathbf{C}$ is known and *vice versa*. ∎

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

Lemma 2. A diagram Δ is a 1-diagram if and only if the vertices of the diagram graph $\imath(\Delta)$ have vertex degree less or equal to 3.

Proof. By definition no vertex of the diagram is contained in more than a single arc. Adding the two edges along the Hamiltonian cycle $\mathcal{H}$, we find the $\deg(v) \leq 3$ for all vertices of the diagram graph. ∎

The diagram graphs of 1-diagrams are closely related to cubic Hamiltonian graphs. The latter are studied in detail in section 9.4 of reference [43].

Definition. Let $\alpha = \{i,j\}$ with $i < j$ be an arc of a diagram. We write $\bar{\alpha} \stackrel{\text{def}}{=} [i,j] \subset \mathbb{R}$ for the associated interval.

Definition. Two arcs of a diagram are *consistent* if they can be drawn without intersection in the same half-plane.

An obvious algebraic characterization is

Lemma 3. 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) $\bar{\alpha} \subseteq \bar{\beta}$.

(iii) $\bar{\beta} \subseteq \bar{\alpha}$.

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

Case (iv) is ruled out by definition in 1-diagrams. 1-diagrams can be used to model the base pairing interactions in nucleic acids. Indeed, the classical definition of a secondary structure [44] requires that each base pairs with at most one other nucleotide. The second defining condition is the absence of *pseudoknots* which can be expressed in terms of the contact matrix in the following from: If $\mathbf{C}_{ij} = \mathbf{C}_{kl} = 1$ and $i < k < j$ then $i < l < j$. That is, if the intervals of two arcs $\{i, j\}$ and $\{k, l\}$ have non-empty intersection then one is contained in the other [30]. Using the above terminology we have the following

Definition. A secondary structure is a 1-diagram in which any two arcs are consistent.

A graph that can be embedded in the plane such that all its vertices lie on the exterior region is called *outerplanar*. This class of graphs was introduced and characterized in terms of subgraphs in ref. [3]. Another interesting characterization in terms of a spectral invariant is discussed in [4].

Lemma 4. A 1-diagram Δ is a secondary structure if and only if its diagram graph $\imath(\Delta)$ is outerplanar.

Proof. A secondary structure is a 1-diagram Δ in which all arcs can be drawn without intersection in the same half-plane. Equivalently, all arcs can be drawn inside the Hamiltonian cycle $\mathcal{H}$ in $\imath(\Delta)$. ∎

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 $s_q =$ '`)`'. Since 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 [15]. 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 [13].

The close resemblance of cubic Hamiltonian graphs [43] and diagram graphs of 1-diagrams suggests to investigate their relation in some more detail. 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 since the triangle $1,2,3$ cannot be obtained from a cubic graph.

Definition. An arc α is an *undisturbed hairpin* if
(i) either there is no $\beta \in \Omega$ for which $\bar{\beta} \subset \bar{\alpha}$ is true, or $\bar{\beta} \subset \bar{\alpha}$ for all $\beta \in \Omega$.
(ii) α is consistent with all $\beta \in \Omega$.

Lemma 5. A diagram graph $\imath(\Delta)$ is homeomorphic from a cubic Hamiltonian graph Γ if and only if the set of arcs Ω is non-empty and Δ does not contain an undisturbed hairpin.

Proof. Consider a subset C of vertices such the induced subgraph in $\imath(\Delta)$ is a cycle C in $\imath(\Delta)$. It is easy to see that C can be contracted to a cycle that contains no vertices of degree 2 (in the original graph) if and only if it contains at least 3 vertices of degree larger than 2. These contractions can be performed independently in all such induced subgraphs, thus $\imath(\Delta)$ is homeomorphic from a cubic Hamiltonian graph if and only if each minimal cycle contains at least three vertices with degree 3. Observing that undisturbed hairpins are exactly the minimal cycles that contain two vertices of degree 3 completes the proof.

Corollary. A diagram graph of a secondary structure is not homeomorphic from a cubic Hamiltonian graph.

2.2. The Inconsistency Graph of a Diagram

Definition. 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 [43], where a result analogous to the following theorem is proved:

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 $\imath(\Delta)$ is planar.
(iii) The inconsistency graph $\Theta(\Delta)$ is bipartite.

Proof. (i $\Longleftrightarrow$ ii) Δ can be drawn without intersection arcs if and only if $\imath(\Delta)$ is planar because the Hamiltonian cycle $\mathcal{H}$ of $\imath(\Delta)$ 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.
(i $\Longleftrightarrow$ iii) If Θ is bipartite then there are two disjoint subsets Ω_U and Ω_L of Ω such that all arcs within the same subset are mutually consistent. Thus all arcs of Δ can be drawn without intersection. Conversely, if we can draw Δ without intersecting arcs then all arcs above and below the x-axis are mutually consistent, i.e., two arcs can be inconsistent only if they lie on different sides of the x-axis. Thus $\Theta(\Delta)$ is bipartite. ∎

Most of the literature on linked diagrams deals with *complete* diagrams, that is, each vertex $x \in [n]$ is incident with an arc [42, 20, 34]. It is straightforward to extend Touchard's definition of reducible diagrams to the incomplete diagrams considered here:

Definition. A diagram $([n], \Omega)$ is *reducible* if there is 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*.

It will be convenient to say that an interval $[r, s]$ *supports* an arc α if $\bar{\alpha} \subseteq [r, s]$. Let $\Omega_{[r,s]}$ be the set of arcs supported by $[r, s]$. We shall say that $([r, s], \Omega_{[r,s]})$ is a sub-diagram of $([n], \Omega)$ if $[r, s]$ fulfils (i). The sub-diagram is non-trivial if $\Omega_{[r,s]}$ is neither empty nor equals Ω. A diagram is therefore reducible if and only if it contains a non-trivial sub-diagram. Let $([p, q], \Omega_{[p,q]})$ and $([r, s], \Omega_{[r,s]})$ be two sub-diagrams of Δ. Then $[p, q] \subset [r, s]$ implies $\Omega_{[p,q]} \subset \Omega_{[r,s]}$ and $([p, q] \cap [r, s], \Omega_{[p,q]} \cap \Omega_{[p,q]})$ is again a sub-diagram of Δ. The sub-diagrams of Δ therefore form a lattice with respect to inclusion.

Lemma 6. A diagram Δ is irreducible if and only if its inconsistency graph $\Theta(\Delta)$ is connected.

Proof. (i) Suppose $([n], \Omega)$ is reducible, and let $([p, q], \Omega_{[p,q]})$ be a non-trivial sub-diagram. Thus there are no arcs that are incident with vertices both in $[p, q]$ and $[n] \setminus [p, q]$ and an arc α supported by $[p, q]$ is consistent with any arc β not supported by $[p, q]$. For all $\beta \in \Omega \setminus \Omega_{[p,q]}$ it is true therefore that β is not connected

with any $\alpha \in \Omega_{[p,q]}$ in $\Theta(\Delta)$ and $\Theta(\Delta)$ decomposes into at least two non-empty components.
Now assume that $\Theta(\Delta)$ is not connected. The *support* of a component Θ' of $\Theta(\Delta)$ in $[n]$ is the union of all intervals $[r, s]$ where r and s are incident with arcs in Θ' and all vertices in $[r+1, s-1]$ are unpaired. Either the support of Θ' is connected, i.e., in which case it forms an interval fulfilling the conditions (i), (ii), and (iii), and Δ is a reducible diagram, or it contains a "hole" $[u, v]$ that contains a vertex x incident with an arc $\gamma \notin \Theta'$. Since γ is consistent with all arcs of Θ', it cannot be incident to any vertex outside $[u, v]$ (otherwise it would need to cross at least one arc of Θ'). Thus $[u, v]$ fulfills conditions (i), (ii), and (iii), and Δ is reducible.
∎

The proof of lemma 6 implies an even stronger result: 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 conform the conventional decomposition into stems and loops, however, which form the basis of the standard energy model of nucleic acid secondary structures [9]. The notion of a stem trivially generalized to arbitrary 1-diagrams:

Definition. 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 $p = 0, \dots, h-1$.

Lemma 7. Let Ψ be a stem in the 1-diagram Δ. Then the arcs of Ψ are either all isolated vertices or they are contained in the same component of the inconsistency graph $\Theta(\Delta)$; all arcs of a stem have the same adjacent vertices in $\Theta(\Delta)$.

Proof. It suffices to show that an arc β that is inconsistent with one arc $\alpha \in \Psi$ must be inconsistent with all arcs of the same stem. We observe that the arcs of Ψ have the form $\alpha_p = \{i+p, j-p\}$ with $p = 0, \dots, h$. Suppose β is inconsistent with α_p for some p. Then it involves a vertex k with $i+p < k < j-p$ and a vertex l with either $l < i+p$ or $l > j-p$. Since there is at most one arc attached to each vertex, β cannot involve the vertices between $i+p$ and $i+h$ or the vertices between $j-h$ and $j-p$ since they are already used by the arcs of the stem Ψ; hence we have $i+h < k < j-h$. A similar argument shows that l satisfies either $l < i$ or $l > j$, and β is therefore inconsistent with all arcs $\alpha_p \in \Psi$. Hence β is adjacent in $\Theta(\Delta)$ to all $\alpha_p \in \Psi$ and thus they belong to the same component of the inconsistency graph. If all arcs in Ω are consistent with (the arcs of) the stem Ψ they appear as isolated vertices in $\Theta(\Delta)$. ∎

2.3. Bi-Secondary Structures

Definition. A *bi-secondary structure* is a 1-diagram that can be drawn 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. Thus $\Omega = \Omega_U \dot\cup \Omega_L$ and the two diagrams $([n], \Omega_U)$ and $([n], \Omega_L)$ are secondary structures. Bi-secondary structures are therefore "superpositions" of two secondary structures.

Theorem 2. Let Δ be a 1-diagram. Then the following statements are equivalent:
(i) Δ is a bi-secondary structure.
(ii) $\imath(\Delta)$ is planar.
(iii) $\Theta(\Delta)$ is bipartite.
(iv) Among any three arcs of Δ at least two are consistent.
(v) $\Theta(\Delta)$ does not contain a triangle.

Proof. The equivalence of (i), (ii), (iii) is established in Theorem 1 for all diagrams. The equivalence of (iv) and (v) follows immediately from the definition of $\Theta(\Delta)$. The implication (iii $\Longrightarrow$ v) is obvious. It remains to show that (iv) or (v) indeed implies (i), (ii), or (iii). We shall prove that $\neg$(ii) implies $\neg$(iv).
Suppose $\imath(\Delta)$ is not planar. Then by Kuratowski's theorem implies that it contains a subgraph S that is homeomorphic from either the complete graph K_5 or the complete bipartite graph $K_{3,3}$ [22]. The vertex degree in $\imath(\Delta)$ is at most 3. Therefore it cannot contain a subdivision S of K_5 which would have has five vertices with vertex degree 4. Any subdivision S of $K_{3,3}$ contains exactly six vertices $1, 2, 3, 4, 5, 6$ of degree 3. Of the three edges incident with these vertices two must belong to the Hamiltonian cycle $\mathcal{H}$ because there is at most one arc of Δ attached to each vertex. Thus S contains a cycle $\mathcal{H}'$, containing the vertices 1 through 6, that corresponds to $\mathcal{H}$ in the sense that all edges of $\mathcal{H}$ that have been inherited by S belong to $\mathcal{H}'$. A remaining edge at 1 through 6 corresponds to an arc of Δ, hence it is directly connected to another vertex of degree 3. Thus $\mathcal{H}'$ is a Hamiltonian arc of S. Without loosing generality we may assume that the vertices 1 through 6 are ordered along $\mathcal{H}'$ and hence also along $\mathcal{H}$. Since S is a subdivision of $K_{3,3}$ the three remaining edges (i.e., arcs of Δ) must be $\{1, 4\}$, $\{2, 5\}$, and $\{3, 6\}$. Obviously, they are mutually inconsistent. Non-planarity of $\imath(\Delta)$ implies therefore the existence of three arcs that are mutually inconsistent. ∎

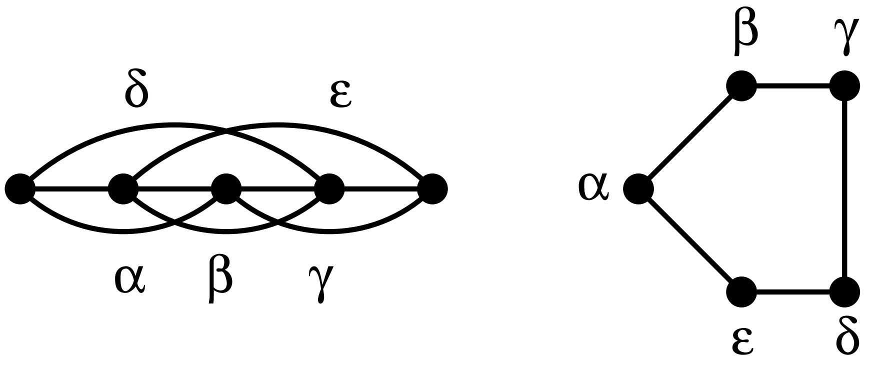


Figure 1: Theorem 2 is not valid for general diagrams. The inconsistency graph of the diagram Δ_5 is a pentagon and hence is neither bipartite nor does it contain a triangle.

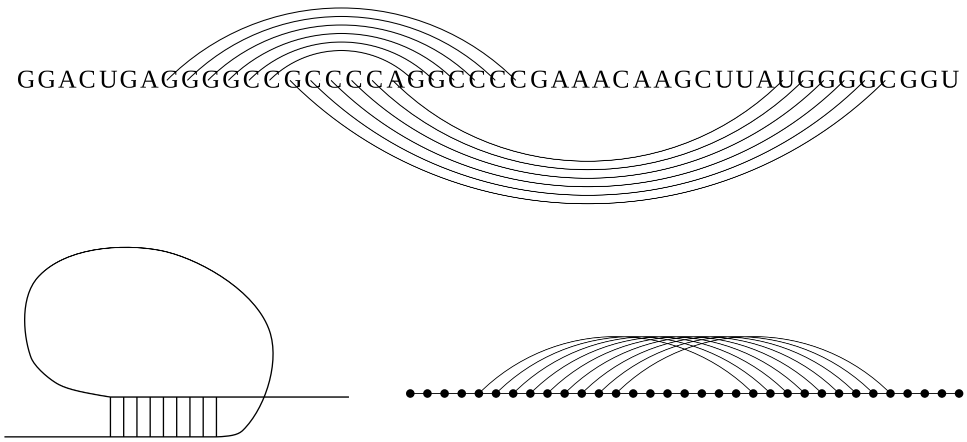


Figure 2: The contact structure of the proposed SRV-1 frameshift signal contains a pseudo-knot, see reference [41]. Pseudoknots such as this one belong to the class of bi-secondary structures.
Knots such as the one in the lower part of the figure do not belong to the class of bi-secondary structures.

The equivalence of (iii) and (v) does not hold for general diagrams. A counterexample is shown in figure 1. The virtue of bi-secondary structures is that they capture a wide variety of RNA pseudoknots, figure 2 (upper part), while at the same time they exclude true knots such as the structure in the lower part of figure 2.

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 $s_q =$ '`]`'.

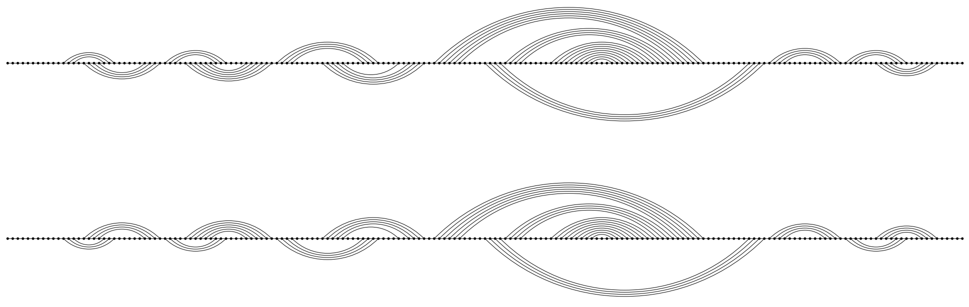


Figure 3: Two diagrams encoding the 3' non-coding region of tobacco mosaic virus RNA [1]. The upper diagram corresponds to the normal form, the lower diagram maximizes the number of upper arcs.

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 since the component is bipartite. Lemma 7 implies that all arcs in a stack are written with the same type of brackets in normal form because they have a common neighboring vertex and hence they all belong to the same class of the partition.

2.4. Beyond Bi-Secondary Structures

A *color partition* of a graph Γ is partition $V = V_1 \cup V_2 \cup \ldots \cup V_c$ 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:

Lemma 8. Let $\Delta = ([n], \Omega)$ be a diagram and let $\mathcal{V} : \Omega = \Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_c$ 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 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:

Corollary. Let Δ be a 1-diagram. Then
(i) Δ is a secondary structure iff $\chi(\Theta(\Delta)) = 1$;
(ii) Δ is a bi-secondary structure iff $\chi(\Theta(\Delta)) \leq 2$.

The chromatic number $\chi(\Theta(\Delta))$ may therefore serve as a measure for the structural complexity of a contact structure.

The following example shows that there are natural RNA structures that have a chromatic number $\chi(\Theta(\Delta)) > 2$. The *Escherichia coli* α-operon mRNA folds into a structure that is required for allosteric control of translational initiation [40]. Compensatory mutations have defined an unusual pseudo-knotted structure [39], the thermodynamics of which were subsequently investigated in detail [10]. The diagram of its contact structure cannot be drawn without intersections, see Figure 4.

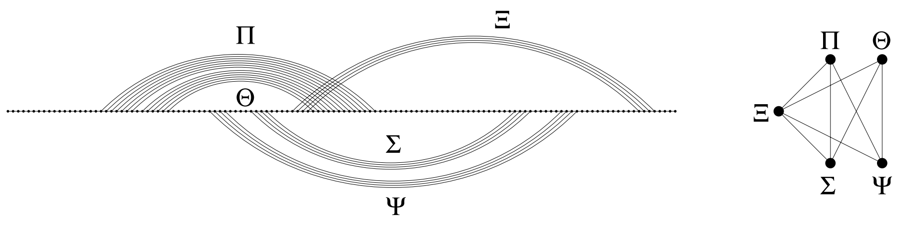


Figure 4: Diagram of the contact structure of E. coli α-mRNA. The structure contains 5 stems, labeled by uppercase Greek letters. As a consequence of lemma 7 we may choose the color partition if $\Theta(\Delta)$ such that all arcs in a stem have the same color. It therefore suffices to draw the inconsistency 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 $\chi(\Theta(\Delta)) = 3$.

Graphs with moderate chromatic numbers can be characterized by results similar to Kuratowski's theorem for planar graphs.

Proposition. [5] Let $k \leq 4$. A graph Γ with chromatic number $\chi(\Gamma) \geq k$ contains a subdivision of the complete graph K_k.

Remark. The generalization of this proposition to $k > 4$ is known as Hajós' conjecture. It is false for $k \geq 7$ and unsolved for $k = 5$ and $k = 6$ [16].

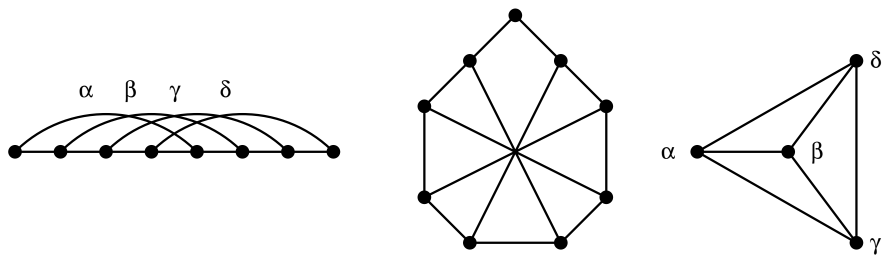


Figure 5: The graph V_8^* and its inconsistency graph.

The graph invariant μ introduced by Colin de Verdière [4] leads to the same hierarchy of structures for small μ:

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

It is tempting therefore to conjecture that $\chi(\Theta(\Delta)) \leq 3$ might correspond to diagram graphs with Colin de Verdière invariant $\mu \leq 4$. The graphs with $\mu \leq 4$ have recently been identified as the *flat* or *linklessly embeddable* graphs [24]. A useful characterization of this class of graphs is proved in [28, 29]: "A graph is non-flat if and only if it has no minor in the so-called Petersen family". The graph V_8^*, figure 5, is a valid diagram graph. It is easy to check that V_8^* is flat and that its inconsistency graph is $\Theta(V_8^*) = K_4$. Hence there are flat diagram graphs for which $\chi(\Theta(\Delta)) \geq 4$. We do not know whether all diagram graphs with $\chi(\Theta(\Delta)) \leq 3$ are flat.

2.5. A Metric for 1-Diagrams

An interesting algebraic interpretation of secondary structures was proposed in [26]. Interpreting each arc $\{i, j\}$ as a transposition (i, j) on $[n]$ we may assign the permutation

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

to each diagram Δ.

Lemma 9. (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.

Proof. (i) Since the arcs of an 1-diagram are disjoint we find only 1-cycles (the unpaired vertices) and 2-cycles (the arcs) in the cycle decomposition of $\pi(S)$. Thus $\pi(S)$ is an involution. The claims (ii) and (iii) are obvious. ∎

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

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

where $\text{cyc}(\pi)$ 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 [26] for a detailed discussion. Since the involutions form a subset of S_n we have

Theorem 3. The function

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

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 [8, 32].

2.6. The Intersection Theorem

The virtue of equ.(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 with pairs of monomers may form a base pair (**AU**, **UA**, **GC**, **CG**, **GU**, **UG** in the case of RNA).

Definition. 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 nonempty.

The proof of this result in ref. [27] is valid for all 1-diagrams, not only for secondary structures. The intersection theorem sets the stage for shape space covering: it allow close-by sequences to fold into structures that are as different as desired — given a suitable folding potential. Further applications of equ.(3) can be found in [48].

Neutral networks in sequence space are modeled as random graphs in [27]. 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, must be determined computationally for a particular choice of the folding potential. Energy model for bisecondary structures and folding of such structures will be discussed elsewhere.

3. Combinatorics

3.1. Enumeration 1-Diagrams

The number X_n of all diagrams on n vertices is $X_n = 2^{(n-1)(n-2)/2}$ since there are $(n-1)(n-2)/2$ possible arcs [33], which can be arbitrarily combined to form a diagram.

In lemma 9 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 [49]:

Proposition. The number T_n of involutions fulfills the recursion

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

and has the asymptotic form

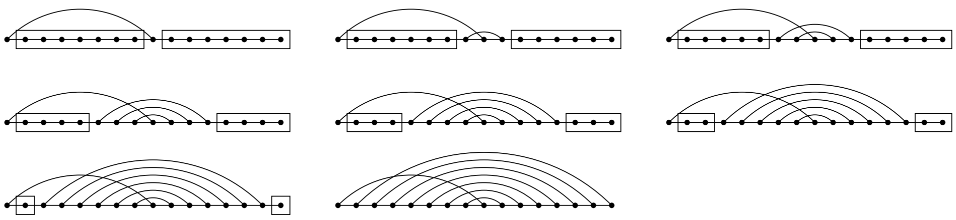
$$T_n \sim \frac{1}{\sqrt{2}} n^{n/2} \exp\left(-\frac{n}{2} + \sqrt{n} - \frac{1}{4}\right)\,.$$

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

Theorem 5. The number of 1-diagrams fulfills the recursion

$$D_{n+2} = D_{n+1} + (n+1)D_n - D_{n-1} + D_{n-2} \quad n \geq 2 \qquad D_0 = D_1 = D_2 = 1,\ D_3 = 2\,.$$

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

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

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

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. 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 $\chi(\Theta(\Delta)) \leq c$. We find

Theorem 6. $D_n(c) \leq (2c+1)^n$.

Proof. Consider a 1-diagram $\Delta = ([n], \Omega)$ with $\chi(\Theta(\Delta)) \leq c$. 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.

3.2. Secondary Structure

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:

$$\begin{aligned} S_{n+1} &= S_n + \sum_{k=m}^{n-1} S_k S_{n-k-1}, \qquad n \geq m+1 \\ S_0 &= S_1 = \ldots = S_{m+1} = 1 \end{aligned} \tag{9}$$

This expression has first been derived by Waterman [44]; 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 [14] for details. Asymptotically, these numbers behave as

$$\Psi_n^{(m,l)} \sim B_{m,l} n^{-3/2} A_{m,l}^n \,. \tag{10}$$

The most important numbers are collected in table 1. A more detailed table can be found in [14].

Table 1. The constants A_{ml} in equ.(10).

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

Detailed combinatorial studies on various aspects of secondary structure graphs are based on equ.(9), see for instance [25, 36, 44, 46, 47, 45, 14]. In the following we shall make use of the number

$$s(n,k) = \frac{1}{k}\binom{n-k}{k+1}\binom{n-k-1}{k-1} \tag{11}$$

of secondary structures of length n with k base pairs. This closed formula was recently derived in [30].

3.3. Bi-Secondary Structure

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

Lemma 10. $D_n(2) \le \max\limits_{\substack{0\le k+l\le n/2\\ l\le k}} \dfrac{n}{2}\dbinom{n-k-1}{k-1}\dbinom{n-k}{k+1}\dbinom{n-2k}{2l}\dbinom{2l}{l}$.

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

Proof. Let $A_n(k,l)$ denote argument of the maximum in lemma 10. It is straight-forward to compute

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

Set $A \stackrel{\text{def}}{=} \max\{\, A(x,y) \,|\, 0 \le x+y \le 1/2 \wedge y \le x \,\}$. Then $\lim \sqrt[n]{D_n(2)} \le \exp(A)$. Solving the optimization problem that defines A is straight-forward. 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. ∎

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 [14] 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

$$A_{ml}^{(\kappa)} \stackrel{\text{def}}{=} \lim_{n\to\infty} \sqrt[n]{\Psi_n^{(m,l;\kappa)}} \tag{12}$$

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

Table 2. Best estimates for the constant $A_{ml}^{(2)}$.
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

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 $5^n \times N_{\text{nf}}/N_{\text{sample}}$, 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 that is described in detail in [38], 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 $\Psi_n^{(m,l)} \times N_{\text{nf}}/N_{\text{sample}}$, 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 [14].

Our best estimates are compiled in table 2. In the biologically interesting case, $m = 3$ and $l = 2$, we find $A_{3,2}^{(2)} \approx 2.35$.

4. 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 pseudoknots are bi-secondary structures (which do not involve nested pseudoknots). Bi-secondary structures correspond to planar graphs while secondary structures form the sub-class of outerplanar graphs.

The inconsistency graph introduced in section 2.2 is a useful construction capturing most of the geometrical features of nucleic acid structure. Indeed, 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 a the sum of the sequence-dependent energy contributions of all loops, see e.g. [9]. 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 structure. While the number of possible contact structures grows faster than exponentially with the lenght 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 sequences, 4^n, exceeds by far the number of possible bi-secondary structures.

The intersection theorem is valid for bi-secondary structures. Given a folding potential similar to the standart energy model for RNA secondary structures, we may expect shape-space covering also for bi-secondary structures. The mathematical results presented in this contribution thus stronly suggest that the incorporation of pseudoknots will not change the qualitative picture of the RNA sequence-structure map as obtained from secondary structures. A conclusive answer, however, will require a computational study using realistic energy parameters for pseudoknots.

References

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

[2] S. Bonhoeffer, J. S. McCaskill, P. F. Stadler, and P. Schuster. RNA multistructure landscapes. a study based on temperature dependent partition functions. *Eur.Biophys.J.*, 22:13–24, 1993.

[3] G. Chartrand and F. Harary. Planar permutation graphs. *Ann. Inst. Henri Poincarè B*, 3:433–438, 1967.

[4] Y. Colin de Verdière. Sur un novel invariant des graphes et un critère de planarité. *J. Comb. Theory B*, 50:11–21, 1990.

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

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

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

[8] 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.

[9] 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.

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

[11] 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. I. Neutral networks. *Monath. Chem.*, 127:355–374, 1996.

[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. Structures of neutral networks and shape space covering. *Monath. Chem.*, 127:375–389, 1996.

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

[14] I. L. Hofacker, P. Schuster, and P. F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 1996. submitted, SFI preprint 94-04-026.

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

[16] D. A. Holton and J. Sheehan. *The Petersen Graph*, volume 7 of *Australian Mathematical Society Lecture Series*. Cambridge University Press, Cambridge, UK, 1993.

[17] W. N. Hsieh. Proportions of irreducible diagrams. *Studies in Appl. Math.*, 52:277–283, 1973.

[18] M. A. Huynen. Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43:165–169, 1996.

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

[20] D. Kleitman. Proportions of irreducible diagrams. *Studies in Appl. Math.*, 49:297–299, 1970.

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

[22] K. Kuratowski. Sur le problème des courbes gauches en topologie. *Fund. Math.*, 15:271–283, 1930.

[23] A. Loria and T. Pan. Domain structure of the ribozyme from eubacterial ribonuclease P. *RNA*, 2:551–563, 1996.

[24] L. Lovász and A. Schrijver. The Colin de Verdière number of linklessly embeddable graphs. preprint, 1996.

[25] R. C. Penner and M. S. Waterman. Spaces of RNA secondary structures. *Adv. Math.*, 101:31–49, 1993.

[26] C. Reidys and P. F. Stadler. Bio-molecular shapes and algebraic structures. *Comp. & Chem.*, 20:85–94, 1996. SFI preprint 95-10-098.

[27] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.

[28] N. Robertson, P. Seymore, and R. Thomas. Petersen family minors. *J. Comb. Theory B*, 64:155–184, 1995.

[29] N. Robertson, P. Seymore, and R. Thomas. Sachs' linkless embedding conjecture. *J. Comb. Theory B*, 64:185–227, 1995.

[30] W. R. Schmitt and M. S. Waterman. Linear trees and RNA secondary structure. *Discr. Appl. Math.*, 12:412–427, 1994.

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

[32] 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.Lond.B*, 255:279–284, 1994.

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

[34] P. R. Stein. On a class of linked diagrams, I. Enumeration. *J. Comb. Theory A*, 24:357–366, 1978.

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

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

[37] M. Tacker, W. Fontana, P. F. Stadler, and P. Schuster. Statistics of RNA melting kinetics. *Eur.Biophys.J.*, 23:29–38, 1994.

[38] M. Tacker, 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, 1996.

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

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

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

[42] J. Touchard. Sur une problème de configurations et sur les fractions continues. *Canad. J. Math.*, 4:2–25, 1952.

[43] K. Wagner and R. Bodendiek. *Graphentheorie II.* B.I. Verlag, Mannheim, Germany, 1990.

[44] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Studies*, 1:167 – 212, 1978.

[45] M. S. Waterman. *Introduction to Computational Biology: Maps, Sequences, and Genomes.* Chapman & Hall, London, 1995.

[46] M. S. Waterman and T. F. Smith. Combinatorics of RNA hairpins and cloverleaves. *Studies Appl. Math.*, 60:91–96, 1978.

[47] M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Math. Biosc.*, 42:257–266, 1978.

[48] J. Weber. *Dynamics on Neutral Evolution.* PhD thesis, Friedrich Schiller University, Jena, January 1997.

[49] H. S. Wilf. *Generatingfunctionology.* Academic Press, San Diego CA, 1994.

[50] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46(4):591–621, 1984.