On the Page Number of RNA Secondary Structures with Pseudoknots

Peter Clote[*] Stefan Dobrev[†] Ivan Dotu[*] Evangelos Kranakis[‡]
Danny Krizanc[§] Jorge Urrutia[¶]

October 1, 2011

Abstract

Let $\mathcal{S}$ denote the set of (possibly noncanonical) base pairs $\{i,j\}$ of an RNA tertiary structure; i.e. $\{i,j\} \in \mathcal{S}$ if there is a hydrogen bond between the ith and jth nucleotide. The *page number* of $\mathcal{S}$, denoted $\pi(\mathcal{S})$, is the minimum number k such that $\mathcal{S}$ can be decomposed into a disjoint union of k secondary structures. Here, we show that computing the page number is NP-complete; we describe an exact computation of page number, using constraint programming, and determine the page number of a collection of RNA tertiary structures, for which the topological genus is known. We describe an approximation algorithm from which it follows that $\omega(\mathcal{S}) \leq \pi(\mathcal{S}) \leq \omega(\mathcal{S}) \cdot \log n$, where the *clique* number of $\mathcal{S}$, $\omega(\mathcal{S})$, denotes the maximum number of base pairs that pairwise cross each other.

1 Introduction

Given an RNA sequence $\mathbf{s} = a_1, \ldots, a_n$, where $a_i \in \{A,C,G,U\}$, a secondary structure S on $\mathbf{s}$ is defined to be a set of unordered pairs $\{i,j\}$ such that:

1. *Watson-Crick or GU wobble pairs:* If $\{i,j\}$ belongs to S, then pair $\{a_i, a_j\}$ must be one of the following canonical base pairs: $\{A,U\}$, $\{U,A\}$, $\{G,C\}$, $\{C,G\}$, $\{G,U\}$, $\{U,G\}$.

2. *Threshold requirement:* If $\{i,j\}$ belongs to S, and $i < j$ then $j - i > \theta$.

3. *Nonexistence of pseudoknots:* If $\{i,j\}$ and $\{k,\ell\}$ belong to S, then it is not the case that $i < k < j < \ell$.

4. *No base triples:* If $\{i,j\}$ and $\{i,k\}$ belong to S, then $j = k$; if $\{i,j\}$ and $\{k,j\}$ belong to S, then $i = k$.

[*] Department of Biology, Boston College, Chestnut Hill, MA 02467, USA. Research supported in part by National Science Foundation NSF grants DMS-1016618 and DMS-0817971, and by Digiteo Foundation.

[†] Institute of Mathematics, Slovak Academy of Sciences, Bratislava, Slovak Republic. Supported in part by VEGA and APVV grants.

[‡] School of Computer Science, Carleton University, K1S 5B6, Ottawa, Ontario, Canada. Research supported in part by NSERC and MITACS grants.

[§] Department of Mathematics, Wesleyan University, Middletown CT 06459, USA.

[¶] Instituto de Matemáticas, Universidad Nacional Autónoma de México.

For steric reasons, following convention, the threshold θ, or minimum number of unpaired bases in a hairpin loop, is taken to be 3 [62, 23]. In contrast, a (general) RNA structure $\mathcal{S}$ on $\mathbf{s}$ is only required to satisfy the following conditions. *(1′)* If $\{i, j\}$ belongs to $\mathcal{S}$, then the ith and jth nucleotide can form a (possibly noncanonical) base pair. *(2)* If $\{i, j\}$ belongs to $\mathcal{S}$, and $i < j$ then $j - i > \theta$. Hence, a (general) RNA structure, comprising the hydrogen bonded nucleotide interactions, may contain pseudoknots and base triples. Throughout the paper, when we refer to *RNA structure*, we mean *general* structure, unless we explicitly mention *secondary structure*.

RNA secondary structure prediction methods generally employ either *(i) thermodynamics*-based dynamic programming approaches, pioneered in Zuker's algorithm [62], as implemented in `mfold` [61], `UNAFold` [34], `RNAfold` [22], `RNAstructure` [35], or *(ii) covariance model* approaches, such as the *stochastic context free grammar* approach implemented in `PFOLD` [29] and `tRNAscan-SE` [32]. The base pair prediction accuracy of thermodynamics-based methods (comparable with covariance model methods) is at most approximately 70% for RNA sequences of at most 700 nt [36]; for a comparative benchmarking of a number of thermodynamics-based and covariance model methods, see the study of Gardner et al. [13]. The most accurate current method of RNA secondary structure prediction uses a hybrid approach, combining the experimental method of selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) with minimum free energy structure prediction using constraints [10]. This hybrid approach yields secondary structure accuracy of approximately 95%, comparable with the manually intensive method of *comparative sequence analysis* [18].

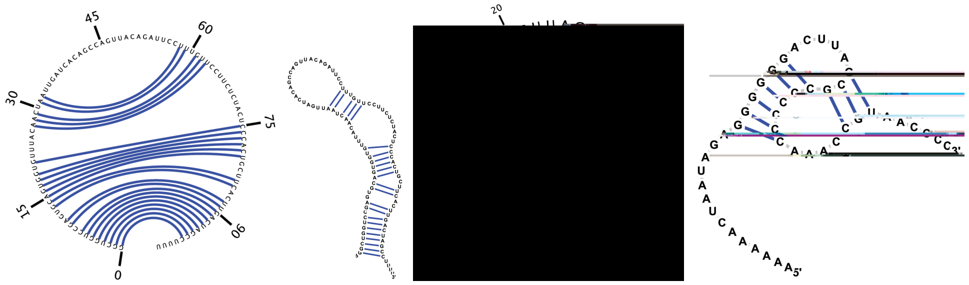

Figure 1: *(a,b)* Pseudoknot-free secondary structure of Y RNA with EMBL access code AAPY01489510/220-119, depicted in panel *(a)* in Feynman circular form, and in panel *(b)* in classical form. *(c,d)* Pseudoknotted structure for the Gag/pro ribosomal frameshift site of mouse mammary tumor virus, depicted in panel *(c)* in Feynman circular form, and in panel *(d)* in classical form. The third and fourth panel of this figure illustrate a type-H pseudoknot (the most common type of pseudoknot), where, for instance, base pair $\{19, 28\}$ crosses base pair $\{25, 40\}$. Images produced with sofware `jViz` [58] from structures taken respectively from Rfam [17] and Pseudobase [53].

The situation is different for *pseudoknotted* structures containing *crossing* base pairs $\{i, j\}$, $\{x, y\}$, such that $i < x < j < y$, where there is a need to improve structure prediction accuracy. Figure 1 illustrates a type-H pseudoknot (the most common type of pseudoknot), where base pair $\{19, 28\}$ crosses base pair $\{25, 40\}$. Indeed, in the case that an RNA structure contains such non-nested base pairs, there is no universally accepted criterion even to define which base pairs form part of the secondary structure and which base pairs form pseudoknots. In fact, given an RNA X-ray structure, different methods surveyed in [47] may yield predictions of different pseudoknotted regions! Another difficulty in pseudoknot structure prediction is the fact that there are no experimentally determined free energies for pseudoknot formation, although Cao and Chen have described a computational method to approximate loop entropies for type-H pseudoknots [7]. Moreover, Lyngsø and Pedersen [33] have shown that minimum free energy pseudoknot structure prediction is an NP-complete problem. This situation is unfortunate, since pseudoknots often play critical biological roles, such as promoting a programmed -1 ribosomal frameshift [51]. For additional biologically important examples of pseudoknot, consult the PSEUDOBASE database of pseudoknots [53, 50].

Pseudoknot prediction algorithms include the genetic algorithm of [1], the maximum weight matching approach of [49], the thermodynamics-based methods of [44, 11, 41, 43] which handle certain subclasses of pseudoknots, the Monte Carlo approaches of [38, 37], heuristics like position specific scoring matrices on tree structures [46] and `ProbKnot` [2], and the exact (exponential-time) methods using tree-width decomposition [60], branch-and-bound [6, 4] and integer programming [40, 45]. Furthermore, using tree-width decompositions, Huang et al. [25] developed fast and accurate genomic search for non-coding RNA pseudoknot structures.

1.1 Preliminaries

A (general) RNA structure uniquely corresponds to a particular *contact map* [55], or *adjacency matrix* [19], $A = (a_{i,j})$, where $a_{i,j} = 1$ if the ith and jth nucleotide form a hydrogen bond, and otherwise $a_{i,j} = 0$. By analyzing the distance and geometry between atoms in the X-ray crystal structure of an RNA molecule, the software `RNAview` [59] determines the collection of hydrogen bonds, including noncanonical bonds [31]. Thus for the purposes of this paper, an RNA structure can be considered as the output of the program `RNAview`.

When depicting both secondary structures and (general) RNA structures, we may add additional edges $\{i, i+1\}$, for $1 \leq i < n$, which correspond to the covalent backbone; however, these edges do not formally belong to the structure. At times we will consider an RNA structure $\mathcal{S}$ to be a collection of base pairs satisfying only conditions $(1', 2)$ given immediately after the definition of a secondary structure in Section 1. At other times, we will variously consider $\mathcal{S}$ to be the corresponding graph just defined, or adjacency matrix, or *circle graph*, which latter is defined in Definition 2. Moreover, since much of the work in this paper concerns the combinatorics of laying out, or decomposing, an RNA structure into a disjoint union of secondary structures, the identity of the nucleotides is not essential, hence will not be mentioned. In particular, we let $\mathcal{S}(n)$ denote the set of all (general) RNA structures on positions (or bases) $1, \ldots, n$.

Definition 1 (Page number) *The* page number *of a structure $\mathcal{S} \in \mathcal{S}(n)$, denoted by $\pi(\mathcal{S})$, is the minimum number n, such that $\mathcal{S}$ can be written as a disjoint union of n secondary structures; i.e. $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_n$, where each $\mathcal{S}_i$ is a secondary structure, and $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for distinct i, j.*

An equivalent graph theoretic formulation of page number is as follows. Let $A = (a_{i,j})$ be the contact map (adjacency matrix) corresponding to a given structure $\mathcal{S}$ for the RNA sequence $\mathbf{s} = s_1, \ldots, s_n$. the *page number* is the minimum number k such that k colors suffice to color all base pairs, with the constraint that if distinct base pairs $\{i, j\}$ and $\{x, y\}$ have the same color, then it is *not* the case that $i \leq x \leq j \leq y$.[*]

[*] Note that Definition 1 requires that base pairs from a triple be placed on separate pages; i.e. base pairs

RNA page k structures generalize the notion of *bisecondary structure*, defined in [20] to be the disjoint union of at most two secondary structures; i.e. page number at most 2. We note that if $\mathcal{S}$ is known to be a bisecondary structure, then by adapting the result of Micali and Vazirani [39], we can describe an algorithm with run time $O(n^{3/2})$ which computes the decomposition of $\mathcal{S}$ into at most 2 pages.

A k-page RNA structure for sequence $\mathbf{s} = s_1, \ldots, s_n$ can be visualized by writing the vertices $1, 2, \ldots, n$ along the spine of a book, where each of the k pages contains a (planar) secondary structure with no crossing edges or base triples. For instance, Figure 2 depicts a 3-page decomposition of an RNA structure.

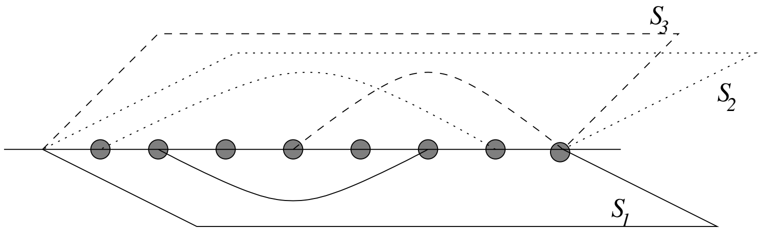

Figure 2: An RNA structure represented in three pages.

Before proceeding, we mention that an obvious greedy algorithm does not necessarily produce the optimal page decomposition for a given RNA structure $\mathcal{S}$. The greedy algorithm proceeds as follows: *(1)* enumerate the base pairs of $\mathcal{S}$ as $b_1, b_2, \ldots, b_m$ and then place b_1 into the first secondary structure $\mathcal{S}_1$. Assume, by induction, that the first i base pairs $b_1, b_2, \ldots, b_i$ have already been placed into disjoint sets $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_j$. Given the next base pair b_{i+1}

1. either for some $r = 1, 2, \ldots, j$ there is no base pair in $\mathcal{S}_r$ that crosses b, in which case we place b_{i+1} into the first such set $\mathcal{S}_r$,

2. otherwise, we create a new secondary structure $\mathcal{S}_{j+1} := \{b_{i+1}\}$ with b_{i+1} as its only element.

In this fashion, secondary structures $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$ are constructed, with $\pi(\mathcal{S}) \leq k$. The following example shows that the greedy algorithm is not optimal. To see this take ten bases numbered $1, 2, \ldots, 10$. Consider the following five base pairs

$$\mathbf{1} = \{7, 10\}, \mathbf{2} = \{3, 6\}, \mathbf{3} = \{1, 5\}, \mathbf{4} = \{2, 8\}, \mathbf{5} = \{4, 9\}$$

and suppose that the base pairs are being considered in this order. The greedy algorithm will output $\mathcal{S}_1 = \{\{7, 10\}, \{3, 6\}\}$. Since $\{1, 5\}$ crosses one of the base pairs of $\mathcal{S}_1$ we must have that $\mathcal{S}_2 = \{\{1, 5\}\}$. Since $\{2, 8\}$ crosses base pairs in both $\mathcal{S}_1$ and $\mathcal{S}_2$ we have that $\mathcal{S}_3 = \{\{2, 8\}\}$. Finally, since $\{4, 9\}$ crosses base pairs in $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_3$ we must have that $\mathcal{S}_4 = \{\{4, 9\}\}$. Therefore the greedy algorithm allocates four pages. Nevertheless three pages are sufficient, namely

$$\mathcal{S}_1 = \{\{7, 10\}, \{1, 5\}\}, \mathcal{S}_2 = \{\{3, 6\}, \{2, 8\}\}, \mathcal{S}_3 = \{\{4, 9\}\},$$

of the form $\{i, j\}, \{i, k\} \in \mathcal{S}$ or $\{i, j\}, \{k, j\} \in \mathcal{S}$ are required to be placed on distinct pages. This replies to a question of one of the anonymous referees.

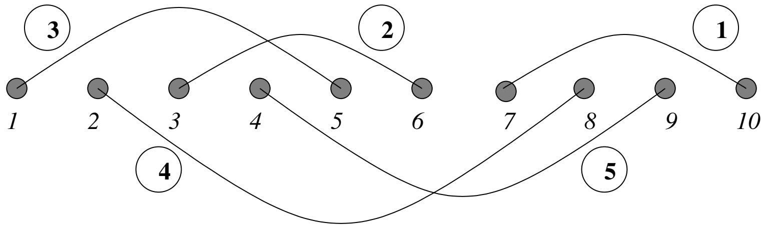

Figure 3: Example of an RNA structure with ten bases and page number three, but for which the greedy algorithm allocates four pages.

and the page number of this RNA structure is three. In the appendix, we show that the greedy algorithm may in fact assign $O(\log n)$ pages for an RNA structure, which only requires a constant number of pages.

RNA structures are related to a class of graphs known in the literature as *circle graphs* [26], which are defined as the intersection graph of the chords of a circle.

Definition 2 (Circle graphs) *Let $I = \{1, \ldots, n\}$ be the index set, considered to constitute a circle consisting of n points, arranged counter-clockwise in order along the periphery of a circle. A circle graph $G = (V, E)$ consists of vertices $v \in V$, which are chords $\{i, j\}$ in the circle, for distinct positions $i, j \in I$, and of undirected edges $e \in E$ formed when two chords overlap; i.e. $e = \{v, w\}$ is of the form $\{\{i, j\}, \{k, \ell\}\}$, where $v = \{i, j\}$ and $w = \{k, \ell\}$ are distinct, and crossing in the sense that where $i \leq k \leq j \leq \ell$, and $i < j$, $k < \ell$.*

Observe that a circle graph G may contain edges of the form $\{i, j\}$, $\{i, k\}$ or of the form $\{i, \ell\}$, $\{k, \ell\}$ or of the form $\{i, j\}$, $\{j, k\}$; in each such case, the edges are considered to be crossing.

To each RNA structure $\mathcal{S}$ we associate its corresponding circle graph $G_{\mathcal{S}}$. It follows that RNA structures can be viewed as constituting a subset of circle graphs. A circle graph G may not correspond to an RNA structure for three reasons.

1. It can happen that there is no way to label positions by nucleotides A,C,G,U such that for each vertex $v = \{i, j\}$ of G, the ith and jth nucleotide can form hydrogen bond. This could happen, if we restrict hydrogen bonds to only canonical (Watson-Crick and wobble) interactions, whenever there are triangles, i.e. a clique of interactions $\{i, j\}$, $\{i, k\}$, $\{j, k\}$ of size 3.

2. There could exist a vertex $v = \{i, j\}$ of G, with $i < j \leq i + 3$.

3. There could exist vertices $u_1 = \{i, j\}$, $u_2 = \{i, k\}$ with $j \neq k$, or vertices $v_1 = \{i, j\}$, $v_2 = \{k, j\}$ with $i \neq k$.

Two base pairs $b = \{i, j\}$ and $b' = \{i', j'\}$, with $i < j$ and $i' < j'$, are said to *cross* if either $i < i' < j < j'$ or $i' < i < j' < j$.

Definition 3 (Chromatic number) *The chromatic number of an RNA structure $\mathcal{S}$, denoted by $\chi(\mathcal{S})$, is defined to be the minimum number n of colors, such that each base pair can be colored in a manner such that crossing base pairs have distinct colors; i.e. n is the* chromatic number $\chi(G_{\mathcal{S}})$ *of the graph $G_{\mathcal{S}}$.*

Clearly, the chromatic number of a structure is the same as the page number $\pi(\mathcal{S})$. In the sequel, we abuse notation and use $\chi(\mathcal{S})$ to denote the chromatic number $\chi(G_{\mathcal{S}})$ of $G_{\mathcal{S}}$.

Definition 4 (Clique number) *If $\mathcal{S}$ is an RNA structure, then let $\omega(\mathcal{S})$ denote the maximum number s of base pairs $b_1, b_2, \ldots, b_s$ in $\mathcal{S}$ such that b_i crosses b_j, for all $i \neq j$.*

Clearly, $\omega(\mathcal{S})$ is the same as the size of the largest *clique* in $G_{\mathcal{S}}$. Figure 4 depicts a structure $\mathcal{S}$ such that $\omega(\mathcal{S}) = 2 < \pi(\mathcal{S}) = 3$.

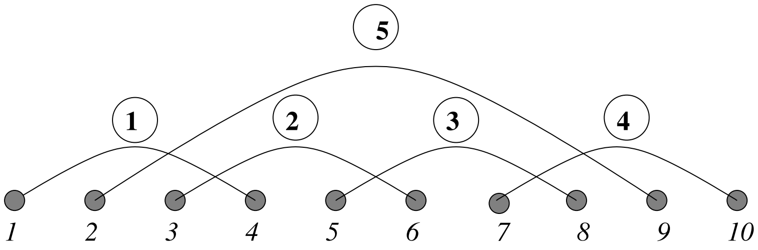

Figure 4: An RNA structure $\mathcal{S}$ on 10 bases with 5 base pairs, having clique number $\omega(\mathcal{S}) = 2$ and page number $\pi(\mathcal{S}) = 3$.

1.2 Related work and outline of paper

In graph theory, a *book embedding* of a graph consists of a linear ordering of the vertices along the "spine of a book" and a planar embedding of its edges on the "pages of the book", i.e. such that no two edges on the same page intersect. The minimum number of pages in which a graph can be embedded is its *page number.* Page number plays a role in circuit design, in the sense that VLSI circuits are created in several layers, or pages [21]. The notion of page number considered in this paper differs from graph theoretic concept of page number, in that the vertices of the graph (in our case the nucleotides) are in fixed positions given by the primary structure. See [9] for additional work on this topic.

Another rigorous classification of pseudoknotted structures uses the notion of topological *genus* g, which in our case corresponds to the the minimum number of handles of a topological surface on which a given RNA structure can be depicted without crossing edges [56, 57].[†] The genus g of a given RNA structure can be computed by a simple application of depth first search, since $g = \frac{P-L}{2}$, where P is the number of base pairs and L the number of closed loops [5]. Vernizzi et al. [56] developed recurrence relations to compute the number of genus g structures of a given RNA sequence, and Bon et al. [5] computed the genus of RNA X-ray structures, using the hydrogen bonding information provided by the program `RNAview` [59]. In [6, 4], Bon and Orland described a novel RNA energy model depending on topological genus, and using this energy model developed a branch-and-bound algorithm to compute the

[†] Yet another classification of pseudoknotted structures involves the notion of k-noncrossing structure, defined by Chen et al. [8], in which no base pair crosses more than k other base pairs. In [24], Huang et al. describe a novel algorithm to compute the minimum free energy, 3-noncrossing, canonical RNA structure, where a structure is defined to be 3-*noncrossing* if it does not contain three or more mutually crossing base pairs, and *canonical*, if each base stack has size at least two. See also Jin and Reidys [27], who analyze the asymptotic number of k-noncrossing structures. Note that *crossing number* is a *local* property, hence easily computable, while *page number* appears to be a *global* property, shown here to be NP-complete.

minimum energy pseudoknotted structure for a given RNA sequence. Recently, Reidys et al. [42] developed a dynamic programming algorithm to compute the minimum free energy structure and partition function over all structures of genus at most 1, where the energy model differs from that of Bon and Orland in several respects, the most important of which concerns the assignment of distinct positive free energies for each of four primitive [4] (irreducible [42]) genus 1 pseudoknots.

As is the case with page number, where the notion used in this paper differs from the standard graph theoretic concept due to the ordering $1, \ldots, n$ of the nodes (nucleotides), there is a difference between the notion of genus of RNA structure [56] and the (general) treatment of genus for unordered graphs. In the latter case, Thomassen [52] has shown that computing the genus is NP-complete, although Filotti et al. [12] have described an algorithm to compute the genus of a (general, unordered) graph in $n^{O(g)}$ time, where n is the number of vertices and g is the genus.

This paper is organized as follows. Section 2 shows that the problem of computing the page number is NP-complete for arbitrary structures with pseudoknots. Faced with NP-completeness, following standard practice in the theory of algorithms, we then give an approximation algorithm to bound the page number of a given structure; in particular, we show that $\omega(\mathcal{S}) \leq \pi(\mathcal{S}) \leq \omega(\mathcal{S}) \cdot \log n$, where $\omega(\mathcal{S})$ denotes the *clique* number of $\mathcal{S}$. In the context of the theory of algorithms, this result is known as a *logarithmic approximation* of page number. Finally, Section 4 describes an exact algorithm to compute the page number and associated layout of base pairs on various pages; in addition, the (optimal) page number is computed for a collection of RNA tertiary structures considered in Bon et al. [5], where the topological genus is computed.

2 NP-completeness of Computing the Page Number

We now show NP-completeness of page number by a polynomial time reduction to the NP-completeness of chromatic number of circle graphs.

Definition 5 (Minimum page number problem) *The minimum page number problem, abbreviated* **MPN***, is the optimization problem of finding the smallest positive integer p such that a given pseudoknotted structure can be represented on p pages.*

More specifically we consider the following decision problem on RNA structures with parameters p (number of pages) and n (number of bases).

MPN(n, p)
Instance: RNA structure $\mathcal{S}$ with pseudoknots on n bases;
positive integer p.
Question: Is $\pi(\mathcal{S}) \leq p$?

We prove the main result of this section by describing a polynomial time transformation of a given circle graph into an RNA structure with the same chromatic number.

Theorem 1 **MPN** *is NP-complete.*

Proof. It is known that computing the chromatic number of circle graphs is NP-complete [14]. A circle graph may not be an RNA structure either because *(i)* there is no labeling of positions $1, \ldots, n$ along the periphery of a circle by A,C,G,U for which the ith and jth nucleotide can form a (possibly noncanonical) base pair for each chord in the circle graph. there is a base pair $\{i,j\}$ with $|j-i| \leq \theta = 3$, or *(ii)* there is a base pair $\{i,j\}$ with $|j-i| \leq \theta = 3$, or since they may have vertices of degree ≥ 2. Finally, it should be mentioned, that although our definition of general RNA structure allows vertices (nucleotides) to have arbitrary degree (e.g. $\{x,y_1\},\{x,y_2\},\ldots,\{x,y_r\}$ for any $0 \leq r \leq n$), there are physical constraints in the three dimensional structure of RNA that effectively prevent arbitrarily large degree r. For any given circle graph $G=(V,E)$ over index set $I=\{1,\ldots,n\}$, we will associate a new circle graph $G'=(V',E')$ over an index set I' of size $2n^2$, in which if chord $\{i,j\} \in V'$, then $|j-i| > \theta = 3$. Moreover, we will ensure that if $\{i,j\}$ and $\{k,\ell\}$ are distinct vertices in V', then i,j,k,ℓ are all distinct. This will allow us to unambiguously assign nucleotides G,C respectively to positions i,j, for each vertex $\{i,j\} \in V'$. It then follows that G' is the circle graph representation $G_{\mathcal{S}}$ for a (general) RNA structure, in which each nucleotide i has a base pair with at most one other nucleotide j, and conditions (1,2,4) are satisfied in the definition of secondary structure given in Section 1. Finally, we show that the chromatic number of G equals that of G'. Hence, if the page number for RNA structures can be computed in polynomial time, then the chromatic number for circle graphs can be computed in polynomial time, contradicting the NP-completeness of chromatic number for circle graphs. Details for the construction of G' and of the chromatic-number preserving embedding $\Phi : G \to G'$ now follow.

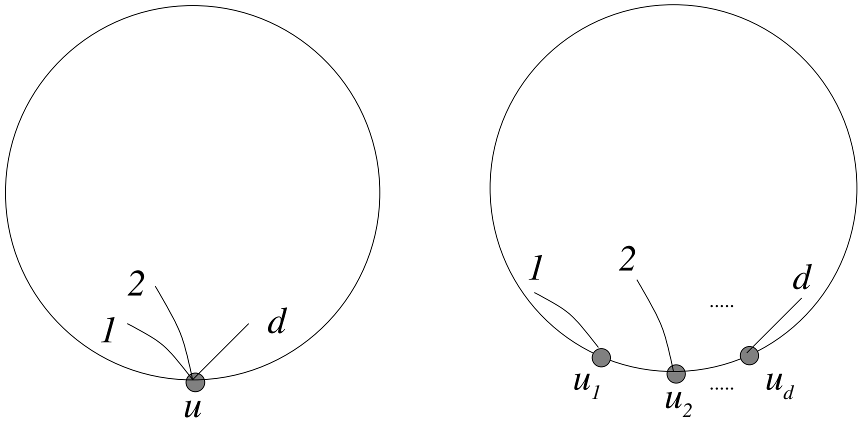

Figure 5: Polynomial time transformation of an arbitrary circle graph G (left panel) into a graph of the form $G_{\mathcal{S}}$ (right panel), for some pseudoknotted RNA structure, such that $\chi(G) = \chi(G_{\mathcal{S}})$.

Let $G=(V,E)$ be a given circle graph, having index set $I=\{1,\ldots,n\}$. We consider the elements of the index set to be laid out in counter-clockwise along the periphery of a circle, in which the vertices of G are chords. Define as follows a new circle graph $G'=(V',E')$ of the same chromatic number.

- **Index set:** Define index set I' of size $2n^2$ by

$$I' = \{(x,i) : x,i \in I\} \cup \{(x,n+i) : x,i \in I\}.$$

The set I' of positions is ordered lexicographically; i.e. $(a,b) < (x,y)$ if either $a < x$ or ($a = x$ and $b < y$). The intent is that to each position $x \in I$, we associate $2n$ positions $(x,1),\ldots,(x,n),(x,n+1),\ldots,(x,n+n)$ in I'.

Define the *one-to-many* association $\Phi : I \to I'$ by

$$\Phi(x) = \begin{cases} (x,1) & \text{if } x \text{ is not incident to any chord of } G \\ (x,y) & \text{if } x \text{ is incident to the chord } \{x,y\}_< \text{ of } G \\ (x,n+y) & \text{if } x \text{ is incident to the chord } \{y,x\}_< \text{ of } G \end{cases}$$

Here, the notation $\{x,y\}_<$ means that $x < y$. Although Φ is *not* a function, since it is many-to-one, we extend Φ to a well-defined injection mapping from vertex set V of G into vertex set V' of G'. By abuse of notation, we will then extend Φ to a well-defined bijection from the set E of edges of G *onto* the set of edges E' of G'.

- **Vertices:** Define $\Phi : V \to V'$ by

$$\Phi(\{x,y\}) = \{\ \{(x,y),(y,n+x)\} \quad \text{if } x < y$$

It follows that $\Phi(\{x,y\}) = \{\Phi(x), \Phi(y)\}$.

- **Edges:** If $\{\{x,y\},\{a,b\}\}$ is an edge of G, then $\{\Phi(\{x,y\}), \Phi(\{a,b\})\}$ is an edge of G'. Edges of G' are defined by crossing chords; i.e. $\{\{u,v\},\{a,b\}\}$ are distinct chords of G, such that $u \le a \le v \le b$ and $u < v$, $a < b$.

See Figure 5 for an example of the transformation just defined. Clearly the transformation $\Phi : G \to G'$ is computable in polynomial time, since the number of edges of G is at most $O(n^2)$. Since the degree of each position $i \in I'$ is at most one, we can assign nucleotide G resp. C to position x resp y for chord $\{x,y\}_<$. It follows that G' is the representation of an RNA structure $G_{\mathcal{S}}$, whose bases correspond to positions $1,\ldots,n'$ along the periphery of the circle corresponding to G'.

We now observe that $G_{\mathcal{S}}$ satisfies conditions 1,2,4 of the definition of a secondary structure, hence is a general RNA structure. Condition 1 is satisfied, since base pairs are all of the form G-C. If $\{x,y\}$ is a vertex in G, with $|y-x| \le \theta = 3$, then $\{(x,y),(y,n+x)\}$ is the image vertex $\{\Phi(x),\Phi(y)\}$ in G', and the positions $(x,y+1),\ldots,(x,n),(x,n+1),\ldots,(x,n+n)$ as well as $(x+1,1),\ldots,(x+1,2n),\ldots,(y,1),\ldots,(y,x-1)$ lie between the image positions $\Phi(x)$ and $\Phi(y)$ of G', hence certainly $|\Phi(y)-\Phi(x)| > 3$. Thus condition 2 is satisfied. Finally, by construction, each position $x \in I'$ has degree at most one, so there are no base triples; condition 4 is satisfied.

We establish a series of claims showing that chord $\{x,y\}$ crosses chord $\{a,b\}$ in G if and only if chord $\{\Phi(x),\Phi(y)\}$ crosses chord $\{\Phi(a),\Phi(b)\}$ in G'. It then follows that the chromatic number of G equals that of G'.

Suppose that chord $\{x,y\}$ crosses chord $\{a,b\}$ in G, so that $x \le a \le y \le b$, and that chords $\{x,y\}$,$\{a,b\}$) are distinct; i.e. this means that $|\{x,y,a,b\}| \ge 3$.

CASE 1: If $x < a < y < b$, then the corresponding images in G' satisfy $(x,y) < (a,b) < (y,n+x) < (b,n+a)$, since the positions I' are ordered lexicographically.

CASE 2: If $\{x,a\},\{x,b\}$ are vertices in G with $x < a < b$, then the corresponding images in G' are $\{(x,a),(a,n+x)\},\{(x,b),(b,n+x)\}$, which satisfy $(x,a) < (x,b) < (a,n+x) < (b,n+x)$.

CASE 3: If $\{a,x\},\{b,x\}$ are vertices in G with $a < b < x$, then the corresponding images in G' are $\{(a,x),(x,n+a)\},\{(b,x),(x,n+b)\}$, which satisfy $(a,x) < (b,x) < (x,n+a) < (x,n+b)$.

CASE 4: If $\{a,x\},\{x,b\}$ are vertices in G with $a < x < b$, then the corresponding images in G' are $\{(a,x),(x,n+a)\},\{(x,b),(b,n+x\}$, which satisfy $(a,x) < (x,b) < (x,n+a) < (b,n+x)$.

It follows that whenever chords $\{x, y\}$ and $\{a, b\}$ cross in G, then chords $\{x, y\}$ and $\{a, b\})$ cross in G'; moreover the crossing in G' is *proper*, in the sense that $\Phi(x) < \Phi(a) < \Phi(y) < \Phi(b)$. Since Φ maps the chords of G onto those of G' in an injective or 1-1 fashion, as shown above, the mapping is in fact an isomorphism, from which it follows that the chromatic number of G equals that of G'; i.e. $\chi(G) = \chi(G')$. This completes the proof of Theorem 1. ■

A related problem concerns the size of the page number of an RNA structure. In [30] it is proved that $\chi(\mathcal{S}) \leq 2^{\omega(\mathcal{S})}$. It is an interesting open question to determine necessary and sufficient conditions which guarantee that the number of pages needed to represent an arbitrary RNA structure can be bounded by a constant independent of the size of the structure. Surprisingly, there is a bound on page number if the circle graph $G_{\mathcal{S}}$ of a given RNA structure $\mathcal{S}$ has no triangles (clique of size 3; if there do not exist three distinct, mutually crossing base pairs in $\mathcal{S}$.

Theorem 2 *Pseudoknotted structures without any triangles can be represented on at most five pages.*

Proof. If the RNA structure $\mathcal{S}$ has no triangles then [28] has shown that $\chi(\mathcal{S}) \leq 8$. This was later improved to $\chi(\mathcal{S}) \leq 5$ by Melnikov (see [26] for additional details). ■

3 Approximation Algorithm for the Page Number

Theorem 1 states that computing the page number of an arbitrary RNA structure is NP-complete. Following standard practice in the theory of algorithms, in this section, we provide an upper bound for page number within a factor of $\log n$.

Theorem 3 *There is a $\log n$ approximation algorithm producing a decomposition of an RNA structure $\mathcal{S}$ on n bases into $O(\pi(\mathcal{S}) \log n)$ pages. Furthermore, the time required to produce the decomposition is $O(n \log n)$.*

Proof. It is clear that a given RNA structure requires at least $\omega(\mathcal{S})$ pages, where $\omega(\mathcal{S})$ is the clique number of $\mathcal{S}$. Thus $\omega(\mathcal{S}) \leq \pi(\mathcal{S})$. We now give a page decomposition algorithm and prove that

$$\pi(\mathcal{S}) \leq \omega(\mathcal{S}) \cdot \log n$$

from which the theorem follows.

The idea of the proof is to use a divide and conquer approach. Look at the bases $\lfloor n/2 \rfloor, \lfloor n/2 \rfloor + 1$ of the structure depicted in Figure 6, where for simplicity, we assume that n is even.

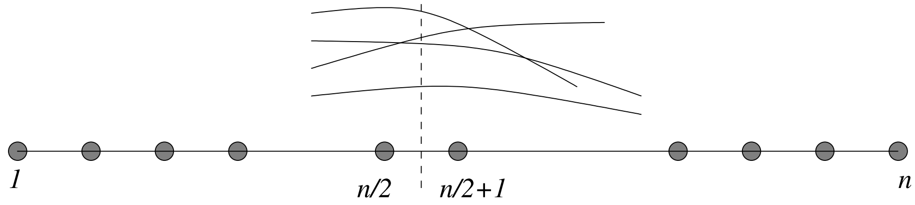

Figure 6: Using divide and conquer to produce a logarithmic approximation for page number of an RNA structure.

Consider all the base pairs $\{i, j\}$ such that $i \leq \lfloor n/2 \rfloor \leq \lfloor n/2 \rfloor + 1 \leq j$. These base pairs can be colored with at most $\omega(\mathcal{S})$ colors, thus resulting in at most $\omega(\mathcal{S})$ pages. This coloring can be found in time $O(n \log n)$. Indeed, the graph induced on the set of edges connecting bases between $1, \ldots, \lfloor n/2 \rfloor$ and $\lfloor n/2 \rfloor + 1, \ldots, n$ has the property that each of its edges crosses the vertical dashed line depicted in Figure 6, and as such is a permutation graph (see [16]). Now remove all the base pairs crossing this vertical line. Two RNA structures result, the first from $1, \ldots, \lfloor n/2 \rfloor$, and the second from $\lfloor n/2 \rfloor + 1, \ldots, n$. Since the base pairs of $1, \ldots, \lfloor n/2 \rfloor$ and $\lfloor n/2 \rfloor + 1, \ldots, n$ do not interfere with each other, the original structure can be colored with the $\omega(\mathcal{S})$ colors required to color the base pairs crossing the vertical line plus the maximum of the number of colors required to color the RNA structures in $1, \ldots, n/2$ and $n/2 + 1, \ldots, n$. If $\chi(n)$ is the minimum number of colors required to color an RNA structure on n bases then it follows that

$$\chi(n) \leq \omega(\mathcal{S}) + \max\{\chi(\lfloor n/2 \rfloor), \chi(n - \lfloor n/2 \rfloor)\}.$$

Applying this technique recursively, we derive that $\pi(\mathcal{S}) \leq \omega(\mathcal{S}) \log n$, as desired.

Concerning the time required to produce the decomposition, observe that each step of the divide and conquer algorithm is required to find a clique in a permutation graph consisting of the base pairs connecting the bases of the structures $1, \ldots, \lfloor n/2 \rfloor$ and $\lfloor n/2 \rfloor + 1, \ldots, n$. Algorithms for finding such a clique are known and require time $O(n \log n)$ (see [16] for more details on algorithms for permutation graphs). ■

4 Exact Computation of Page Number

Here we present an algorithm using *Constraint Programming* [54] to find the page number for a given RNA tertiary structure. Our approach is divided into three different steps:

1. Given a set $\mathcal{S}$ of (possibly noncanonical) base pairs $\{i, j\}$, we collapse $\mathcal{S}$ into a set H of helices $\{h_i\}$, where each h_i is represented by the closing base pair $\{h_i^\ell, h_i^r\}$, where h_i^ℓ is the $5'$ or *left* position, which is paired with h_i^r, the $3'$ or *right* position.

2. We generate a graph $G = (V, E)$ in which the set of vertices V is equal to the set of helices H from the previous step, and in which the set of edges is $E = \{(h_i, h_j) : h_i^\ell < h_j^\ell < h_i^r < h_j^r\}$; i.e. there is an edge between crossing base pairs (h_i^ℓ, h_i^r) and (h_j^ℓ, h_j^r).

3. In this final step we solve the minimum vertex coloring problem on G using Constraint Programming.

For the sake of speed we merged the first two steps in our implementation. The details of these phases are given in the following.

4.1 Collapsing helices

Given a set $\mathcal{S}$ of (possibly noncanonical) base pairs $\{i, j\}$ from the tertiary structure of some RNA, as could be computed using the program `RNAview` [59], we collapse base pairs into helices and create a graph G in which each vertex is a helix and each edge represents a pseudoknot between two helices. Formally, given two helices h_i and h_j characterized by their closing base pairs $\{h_i^\ell, h_i^r\}$ and $\{h_j^\ell, h_j^r\}$, we say they represent a pseudoknot if and only if:

$$h_i^\ell < h_j^\ell < h_i^r < h_j^r.$$

```
collapseAndConstructGraph($\mathcal{S}$)
    $\mathcal{S}^* =$ list of $\mathcal{S}$ in lexicographic order
    $V = \emptyset$
    $E = \emptyset$
    $n = |\mathcal{S}^*|$
    lastbp = $\mathcal{S}^*[0]$  first element of $\mathcal{S}^*$
    forall $bp \in \mathcal{S}^* - \{\text{lastbp}\}$
        if lastbp, bp are not consecutive
            $V = V \cup \{\text{lastbp}\}$
            forall $h \in V$
                if pseudoknot(h,bp)
                    $E = E \cup \{h, h(bp)\}$
        lastbp = bp
    $G = (V, E)$
    return $G$
```

Figure 7: Algorithm to collapse base pairs and construct graph.

Figure 7 depicts the algorithm for collapsing the base pairs and constructing the graph G. It scans the list of base pairs in lexicographic order, creating a new vertex/helix each time a base pair is not *consecutive* with respect to the previous one, where two base pairs $\{i, j\}$ and $\{x, y\}$, with $i < x < y < j$ are defined to be consecutive if, and only if:

$$x \in \{i+1, i+2\} \wedge y \in \{j-1, j-2\}.$$

In other words, two base pairs are consecutive if they are either stacked, or separated by a bulge of size 1 or an internal loop of size 2. The algorithm selects the lexicographically least base pair from each helix, where a helix (or stem) is a maximal collection of consecutive base pairs. After a new helix is "closed", we check with all other previously "closed" helices to determine if they form a pseudoknot with the the one we have just scanned. If so, we create an edge between that helix and the one we are just "opening", denoted by $h(bp)$ in the pseudocode.

As can be seen, this algorithm has a worst time complexity of $O(m^2)$ where m is the number of base pairs in $\mathcal{S}$.

4.2 Solving the minimum vertex coloring using CP

Constraint Programming Constraint programming is one the main methodologies for solving hard combinatorial optimization problems. The salient features of CP are its rich modeling language and its computational model based on branch and prune. At the modeling level, CP models a complex application in terms of decision variables, domains which specify the possible values for the variables, and constraints which capture its combinatorial substructures, giving the underlying solver significant information on the application structure.

The computational model of constraint programming is branch and prune. Constraints are used to filter the variable domains by removing values that cannot appear in any soluion. In fact, each constraint is associated with two algorithms: (1) a feasibility algorithm which determines if a constraint can be satisfied in isolation given the current variable domains; (2) a filtering algorithm that removes values from the variable domains that cannot satisfy the constraints given the current domains.

Graph Coloring In graph theory, *vertex coloring* of an input graph G [15] is a special case of graph labeling in which labels, traditionally called "colors", are assigned to the vertices of G, in a manner such that no two adjacent vertices share the same color. The *minimum vertex coloring* problem is to determine, given an input graph G, the minimum number of colors necessary for a vertex coloring of G. The vertex coloring problem is well-known to be NP-complete [15]. Coloring the graph constructed in the previous subsection with a minimum number of colors is of course equivalent to finding the page number of a given RNA tertiary structure. The number of colors used to color the graph is indeed the page number.

CP model To solve the minimum coloring problem, we use a traditional CP formulation which consists of the following components:

- *Variables:* There is a variable c_i for each vertex v_i in the graph, which represents the color to be assigned to that vertex. There is a variable k, which represents the maximum number of colors used (colors are coded as integers so that calculations of minimum and maximum are possible).

- *Domains:* Letting V represent the set of vertices of the graph G constructed above, we define the domains for all variables c_i to be $D = \{0, \ldots, |V| - 1\}$, and the domain for k to be $D(k) = \{1, \ldots, |V|\}$.

- *Constraints:* There is a constraint for each edge in E such that, for an edge between vertices v_i and v_j, there is a constraint $c_i \neq c_j$. There is also a constraint for each vertex v_i of the form $c_i < k$, that ensures that k is greater than any "color" used.

- *Objective:* minimize the number of colors used, i.e. minimize k.

4.3 Results

In [5], Bon et al. defined the notion of topological *genus* of a pseudoknot, and classified a number of RNA tertiary structures from the Protein Data Bank [3] according to genus. In this section we present comparable results with respect to page number for the same RNA tertiary structures considered in [5]. The results from Table 1 were obtained by our implementation of the above-described Constraint Programming algorithm in the COMET programming language [54].

Given an RNA tertiary structure in PDB file format, our program computes the list of lexicographic least base pairs of each helix belonging to a single page, for pages 0,1,2, etc. and then the optimal page number structure is displayed, using different types of bracket (parentheses, square brackets, curly brackets, etc.) for distinct pages. For instance, given the PDB file 6TNA [48] for the 76 nt yeast phenylalanine transfer RNA, we have the following page layout:

```
0 :  (6,65),(18,55),(29,39)
1 :  (12,21),(52,60)
```

```
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
(((((((..[[[[.....(..]]]].((((.........)))).....[[[[[..)....]]]]])))))))....
```

Notice that there is a single pseudoknot at (19,56), between the D-loop and the TΨC-loop. Rather than perform a page layout where base pair $(19, 56)$ is depicted by a square-bracket, with all other base pairs depicted by round-brackets, our program output the previously displayed, equivalent form. Using `jViz` [58], the 2-page pseudoknotted structure of 6TNA, computed by our program is displayed in Figure 8. Although the determination that 6TNA has page number 2 is trivial, this can hardly be said of the 2922 nt sequence of 23S ribosomal RNA with PDB code 1KC8:A (A chain of 1KC8), depicted in Figure 9. As mentioned in Table 1, the only page 4 structure we found in the collection studied by Bon et al. was the chain A of the file with PDB accession code 1KC8, corresponding to the 2922 nt sequence of 23S ribosomal RNA. The (optimal) page number structure for this and all other RNAs appearing in Table 1 can be found at our web server. Each computation took less than one second.

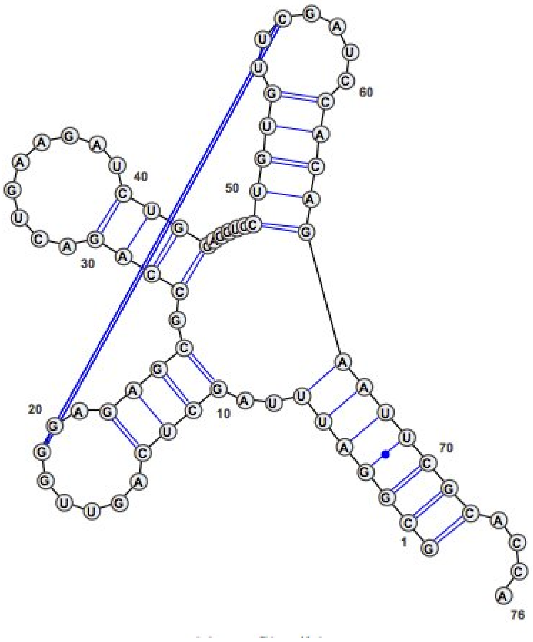

Figure 8: Pseudoknotted structure of the 76 nt yeast phenylalanine transfer RNA with PDB code 6TNA [48]. This example is trivial, since there is only one pseudoknot. Image produced using `jViz` [58].

5 Conclusion and Discussion

In this paper, we have proven that that computing the page number $\pi(\mathcal{S})$ of general pseudoknotted RNA structures is NP-complete. We have observed that an explicit ≤ 2-page layout for structure $\mathcal{S}$ can be determined in time $O(n^{3/2})$, provided that it is known that $\pi(\mathcal{S}) \leq 2$; i.e. $\mathcal{S}$ is a bisecondary structure. Since it is standard policy in the theory of algorithm to

Page Number	PDB file
2	6tna, 4tra, 4tna, 437d, 2tra, 2tpk, 2g1w, 2fk6, 2csx-C, 2a43, 2a2e, 1znn, 1ymo, 1yl4-A, 1yg3, 1yfg, 1y27 1y26, 1x8w, 1voz, 1voy-B, 1vox, 1vov, 1vou-B, 1vos, 1voq, 1vc7, 1vc6, 1vc5, 1vc0, 1vbz, 1vby, 1vbx, 1u8d, 1u6b-B, 1ttt-D, 1tra, 1tn2, 1sz1-E, 1sjf, 1sj4, 1sj3, 1ser, 1qu3, 1qu2, 1qtq, 1qru, 1qrt, 1qrs, 1qf6, 1pnx, 1o0c, 1o0b, 1n77-C, 1n36, 1n34, 1mzp, 1mj1-D, 1l3d, 1l2x, 1kpz, 1kpy, 1kpd, 1jgq-D, 1jgp-D, 1jgo-D, 1j1u, 1il2-C, 1i9v, 1i97, 1i95, 1gtr, 1grz-A, 1gix-C, 1gix-B, 1g59-B, 1fka, 1fir, 1fg0, 1ffz, 1ffy, 1fcw-A, 1f7v, 1f7u, 1exd, 1euy, 1eiy, 1ehz, 1drz, 1cx0, 1c2w, 1c0a, 1asz-S, 1asz-R, 1asy-S, 1asy-R, 1b23
3	2d3o, 2awb-B, 2aw7, 2aw4-B, 2avy, 2aar, 2a64, 1yl3-A, 1yjw-0, 1yjn-0, 1yj9-0, 1yit-0, 1yij-0, 1yi2-0, 1yhq-0, 1y69-0, 1y0q, 1xnr-A, 1xnq-A, 1xmq-A, 1xmo-A, 1xbp-0, 1vqp-0, 1vqo-0, 1vqn-0, 1vqm-0, 1vql-0, 1vqk-0, 1vq9-0, 1vq8-0, 1vq7-0, 1vq6-0, 1vq5-0, 1vq4-0, 1vp0, 1vow-B, 1sm1-0, 1s72-0, 1s1i-3, 1s1h, 1qvg-0, 1qvf-0, 1q86-A, 1q82-A, 1q81-A, 1q7y-A, 1pny-0, 1pnu-0, 1pns-A, 1p9x, 1ond, 1nwy-0, 1nwx-0, 1nkw-0, 1njp-0, 1njo, 1njn, 1njm-0, 1nji-A, 1n8r-A, 1n33-A, 1n32-A, 1m90-A, 1m1k-A, 1kqs-0, 1k9m-A, 1k8a-A, 1k73-A, 1kd1-A, 1k01, 1jzz, 1zy, 1jzx, 1jj2-0, 1j5e, 1j5a, 1ibm-A, 1ibl, 1ibk, 1i96, 1i94, 1hr0, 1hnz, 1hnx, 1hnw, 1fjg, 1ffk-0, 1et4-A, 1ddy-A
4	1kc8-A

Table 1: Page number of all RNA tertiary structures, for which Bon et al. [5] computed the topological genus. In each case, the exact page number was computed by our Constraint Programming algorithm within one second.

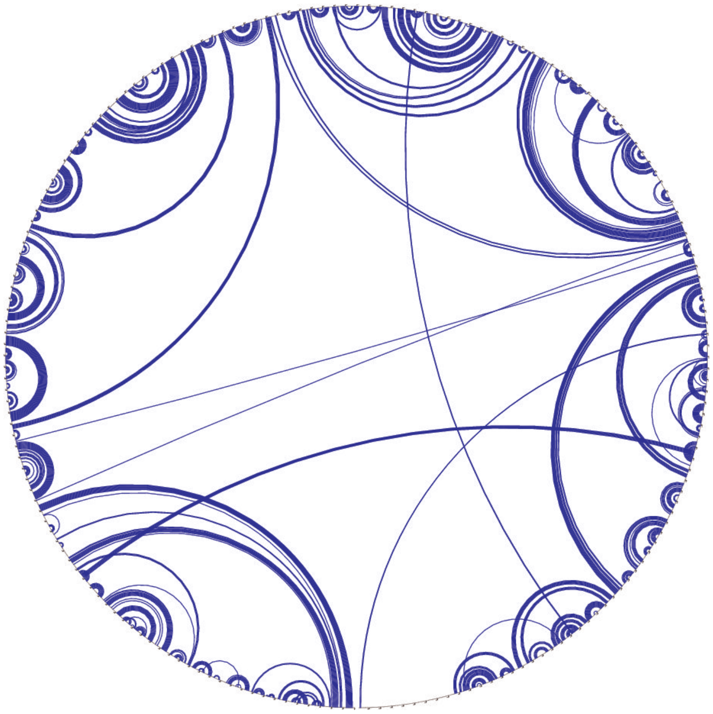

Figure 9: Pseudoknotted structure of the 2922 nt sequence of 23S ribosomal RNA with PDB code 1KC8:A (A chain of 1KC8). Minimum page number layout produced by the Constraint Programming method described in this section; image produced using `jViz` [58].

provide an *approximation* algorithm when faced with NP-completeness, we have a logarithmic approximation, implying that $\omega(\mathcal{S}) \leq \pi(\mathcal{S}) \leq \omega(\mathcal{S}) \cdot \log n$, where the *clique* number of $\mathcal{S}$, $\omega(\mathcal{S})$, denotes the maximum number of base pairs that pairwise cross each other. Finally, we have described a constraint program (CP) which computes the page number of any given RNA structure. Though the CP runs in exponential time, since it implements a branch-and-prune technique, the optimization programming language COMET [54] used in our implementation is extremely efficient, and the page number can be computed for RNA structures from the Protein Data Bank [3] within one second.

The application of topological genus was introduced in the context of RNA structure in [56, 57], and subsequently used in [5] to classify tertiary RNA structures from the Protein Database [3]. In [6, 4], Bon and Orland introduced an energy model for RNA structures, which includes a pseudoknot penalty according to genus, and computed the minimum free energy (MFE) pseudoknotted structure (of arbitary genus, or of genus bound by a user-specified input) using a branch-and-bound algorithm. Reidys et al. [42] gave a dynamic programming algorithm to compute the MFE structure of genus at most 1, where the energy model involved penalties for each of four specific primitive [4] (irreducible [42]) genus 1 pseudoknots.

Since genus is a topological notion, that does not take into account stericity or other molecular constraints, one might consider whether page number provides a better classification of RNA structures. Unfortunately, results from this paper suggest that no reasonable energy model involving page number may exist, since page number is an NP-complete problem.[‡] Table 1 shows that very few biologically relevant structural RNA molecules have large page number, at least among those whose structure is deposited in the Protein Data Bank. It is natural to wonder whether *random* pseudoknotted structures have higher page number. In the Erdös-Renyi model of *random* graphs, where there is an edge between distinct vertices i, j with fixed probability p, we can show[§] that the expected page number of a random pseudoknotted structure $\mathcal{S}_n$ on nucleotide positions $1, \ldots, n$ is *large* and roughly close $\sqrt{n}$; i.e. $E[\pi(S_n)] \in \Omega(\sqrt{n/\log n})$ and $E[\pi(S_n)] \in O(\sqrt{n}\log n)$. Comparing the page number of biologically functional RNA molecules with that of random pseudoknotted structures, we infer that the Erdös-Renyi model of random graph is an *unrealistic* model for random pseudoknotted structures. This situation is similar to the realization made in systems biology, where it is understood that the small-world graphs arising from metabolic pathways and protein-protein interaction networks are different in nature from Erdös-Renyi random graphs.

Acknowledgements

Funding for the research of P. Clote and I. Dotu was provided by the National Science Foundation with grants DMS-1016618 and DMS-0817971, with additional funding to P.C. by Digiteo Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Funding for the research of E. Kranakis was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Mathematics of Information Technology and Complex Systems (MITACS). J. Urrutia was supported in part by CONACYT grant.

[‡] In contrast, see [24] for an energy model concerning crossing number.

[§] Following the suggestion of anonymous referees, we have removed the statement and proof of our analysis of page number of random structures. We plan to pursue and expand the results on page number of random structures in future work, since the results very clearly depend on the notion of randomness used.

Appendix

The following example shows that the greedy algorithm may assign $O(\log n)$ pages for an RNA structure, which only requires a constant number of pages. The sets of base pairs are constructed recursively in stages $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_n$. Initially, $\mathcal{S}_0$ consists of one base pair. Assume that $\mathcal{S}_n$ has been constructed. Then the set S_{n+1} is constructed by appending a copy of $\mathcal{S}_n$ to itself plus the addition of a base pair which 1) surrounds the second copy of $\mathcal{S}_n$, 2) crosses the first copy of $\mathcal{S}_n$, 3) its leftmost base is enclosed inside the rightmost innermost base pair of the first copy of $\mathcal{S}_n$. The resulting structure is depicted in Figure 6 for $n = 0, 1, 2$.

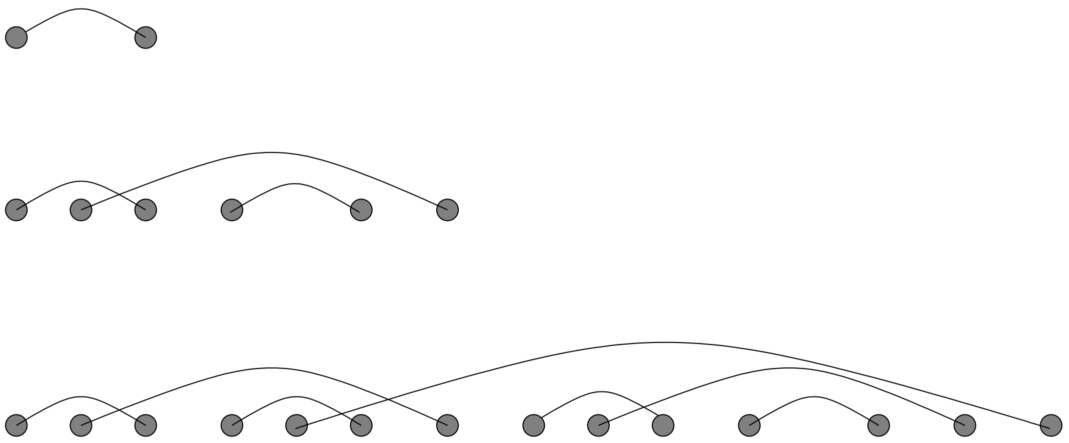

Figure 10: Example of the sequence of sets of base pairs $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2$.

The number B_n of base pairs introduced by the nth step satisfies the recursion $B_n = 2B_{n-1} + 1$ with initial condition $B_0 = 1$. Solving the recursion we see that $B_n = 2^{n+1} - 1$. Now we look at the number of pages being used by the greedy algorithm. Enumerate the base pairs by occurrence of their leftmost base, say $b_0, b_1, b_2, \ldots$. At the kth step, a new page is introduced only if the base pair b_k crosses a previously placed base pair in each of the pages introduced so far; in this case a new page is introduced on which b_k is placed. It is easy to see that the optimal algorithm requires n pages and so does the greedy algorithm.

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. Bellaousov and D. H. Mathews. Probknot: fast prediction of RNA secondary structure including pseudoknots. *RNA.*, 16(10):1870–1880, October 2010.

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Researches*, 28(1):235–242, 2000.

[4] M. Bon and H. Orland. TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Res.*, 0(O):O, May 2011.

[5] M. Bon, G. Vernizzi, H. Orland, and A. Zee. Topological classification of RNA structures. *J. Mol. Biol.*, 379(4):900–911, June 2008.

[6] Michael Bon. *Prédiction de structures secondaires d'ARN avec pseudo-noeuds.* PhD thesis, Ecole Polytechnique, 2009. Ph.D. dissertation in Physics.

[7] S. Cao and S. J. Chen. Predicting RNA pseudoknot folding thermodynamics. *Nucleic. Acids. Res.*, 34(9):2634–2652, 2006.

[8] W. Y. Chen, H. S. Han, and C. M. Reidys. Random K-noncrossing RNA structures. *Proc. Natl. Acad. Sci. U.S.A.*, 106(52):22061–22066, December 2009.

[9] F.R.K. Chung, F.T. Leighton, and A.L. Rosenberg. Embedding graphs in books: A layout problem with applications to VLSI design. *SIAM J. Algebraic Discrete Methods.*, 8(1):33–58, 1987.

[10] K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, 106(1):97–102, January 2009.

[11] R.M. Dirks and N.A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem, 24(13):1664-1677, 2003*, 24(13):1664–1677, 2003.

[12] I. S. Filotti, G. L. Miller, and J. H. Reif. On determining the genus of a graph in $O(\nu^{O}(g))$ steps. In *STOC*, pages 27–37. ACM, 1979.

[13] P. P. Gardner and R. Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC. Bioinformatics*, 5:140, September 2004.

[14] M. R. Garey, D. S. Johnson, G. L. Miller, and C. H. Papadimitriou. The complexity of coloring circular arcs and chords. *SIAM Journal on Algebraic and Discrete Methods*, pages 216–227, 1980.

[15] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman & Co., 1990. New York.

[16] M.C. Golumbic. *Algorithmic graph theory and perfect graphs.* North-Holland, 2004.

[17] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S.R. Eddy. Rfam: an RNA family database. *Nucleic Acids Res.*, 31(1):439–441, 2003.

[18] R. Gutell, J. Lee, and J. Cannone. The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology*, 12:301–310, 2005.

[19] F. Harary. *Graph Theory.* Addison-Wesley, 1994.

[20] Christian Haslinger and Peter F. Stadler. Rna structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bulletin of Mathematical Biology*, 61(3):437–467, May 1999.

[21] Lenwood S. Heath and Sorin Istrail. The pagenumber of genus g graphs is O(g). In *STOC '87: Proceedings of the 19th annual ACM symposium on Theory of computing*, pages 388–397, New York, NY, USA, 1987. ACM.

[22] I. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–3431, 2003.

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

[24] F. W. Huang, W. W. Peng, and C. M. Reidys. Folding 3-noncrossing RNA pseudoknot structures. *J. Comput. Biol.*, 16(11):1549–1575, November 2009.

[25] Z. Huang, Y. Wu, J. Robertson, L. Feng, R. L. Malmberg, and L. Cai. Fast and accurate search for non-coding RNA pseudoknot structures in genomes. *Bioinformatics*, 24(20):2281–2287, October 2008.

[26] T. R. Jensen and B. Toft. *Graph Coloring Problems.* John Wiley and Sons, 1995.

[27] E.Y. Jin and C.M. Reidys. On the decomposition of k-noncrossing RNA structures. *Advances in Applied Mathematics*, 44(1):53–70, 2010.

[28] I. A. Karapetyan. Coloring of arc graphs (in Russian). *Akad. Nauk Armyam. SSR Doklady*, 70:306–311, 1980.

[29] Bjarne Knudsen and Jotun Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, 2003.

[30] A. Kostochka and J. Kratochvil. Covering and coloring polygon-circle graphs. *Discrete Mathematics*, 163(1):299–305, 1997.

[31] N. B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA.*, 7(4):499–512, April 2001.

[32] T. Lowe and S. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleid Acids Research*, 25(5):955–964, 1997.

[33] R. B. Lyngso and C. N. Pedersen. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.*, 7(3-4):409–427, 2000.

[34] N. R. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, 453:3–31, 2008.

[35] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, 101:7287–7292, 2004.

[36] D.H. Mathews, J. Sabina, M. Zuker, and H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[37] D. Metzler and M. E. Nebel. Predicting RNA secondary structures with pseudoknots by MCMC sampling. *J. Math. Biol.*, 56(1-2):161–181, January 2008.

[38] I. M. Meyer and I. Miklos. Simulfold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS. Comput. Biol.*, 3(8):e149, August 2007.

[39] S. Micali and V.V. Vazirani. An $O(\sqrt{|V|}|E|)$ algoithm for finding maximum matching in general graphs. In *Foundations of Computer Science, 1980., 21st Annual Symposium on*, pages 17–27, 1980.

[40] U. Poolsap, Y. Kato, and T. Akutsu. Prediction of RNA secondary structure with pseudoknots using integer programming. *BMC. Bioinformatics*, 10:S38, 2009.

[41] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC. Bioinformatics*, 5:104, August 2004.

[42] C. M. Reidys, F. W. Huang, J. E. Andersen, R. C. Penner, P. F. Stadler, and M. E. Nebel. Topology and prediction of RNA pseudoknots. *Bioinformatics*, 27(8):1076–1085, April 2011.

[43] J. Ren, B. Rastegari, A. Condon, and H. H. Hoos. Hotknots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA.*, 11(10):1494–1504, October 2005.

[44] E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[45] K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):i85–i93, July 2011.

[46] K. Sato, K. Morita, and Y. Sakakibara. PSSMTS: position specific scoring matrices on tree structures. *J. Math. Biol.*, 56(1-2):201–214, January 2008.

[47] S. Smit, K. Rother, J. Heringa, and R. Knight. From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA.*, 14(3):410–416, March 2008.

[48] J. L. Sussman, S. R. Holbrook, R. W. Warrant, G. M. Church, and S. H. Kim. Crystal structure of yeast phenylalanine transfer RNA. I. Crystallographic refinement. *J. Mol. Biol.*, 123(4):607–630, August 1978.

[49] J.E. Tabaska, R.E. Cary, H.N. Gabow, and G.D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14:691–699, 1998.

[50] M. Taufer, A. Licon, R. Araiza, D. Mireles, F. H. Van Batenburg, A. P. Gultyaev, and M. Y. Leung. Pseudobase^{++}: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. *Nucleic. Acids. Res.*, 37(Database):D127–D135, January 2009.

[51] C. A. Theimer and D. P. Giedroc. Equilibrium unfolding pathway of an H-type RNA pseudoknot which promotes programmed -1 ribosomal frameshifting. *J. Mol. Biol.*, 289(5):1283–1299, June 1999.

[52] C. Thomassen. The graph genus problem is NP-complete. *Journal of Algorithms*, 10(4):568–576, December 1989.

[53] F. H. Van Batenburg, A. P. Gultyaev, and C. W. Pleij. Pseudobase: structural information on RNA pseudoknots. *Nucleic. Acids. Res.*, 29(1):194–195, January 2001.

[54] P. Van Hentenryck. *Constraint Satisfaction in Logic Programming.* The MIT Press, Cambridge, MA, 1989.

[55] M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2:295–306, 1997.

[56] G. Vernizzi, H. Orland, and A. Zee. Enumeration of RNA structures by matrix models. *Phys. Rev. Lett.*, 94(16):168103, April 2005.

[57] G. Vernizzi, P. Ribeca, H. Orland, and A. Zee. Topology of pseudoknotted homopolymers. *Phys. Rev. E*, 73(3):031902, March 2006.

[58] K. C. Wiese, E. Glen, and A. Vasudevan. JViz.Rna–a Java tool for RNA secondary structure visualization. *IEEE. Trans. Nanobioscience.*, 4(3):212–218, September 2005.

[59] H. Yang, F. Jossinet, N. Leontis, L. Chen, J. Westbrook, H.M. Berman, and E. Westhof. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, 31(13):3450–3560, 2003.

[60] J. Zhao, R. L. Malmberg, and L. Cai. Rapid ab initio prediction of RNA pseudoknots via graph tree decomposition. *J. Math. Biol.*, 56(1-2):145–159, January 2008.

[61] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.

[62] M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148, 1981.