On the Page Number of Secondary Structures with Pseudoknots

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

March 20, 2011

Abstract

Let S denote the set of (possibly noncanonical) base pairs $\{i,j\}$ of an RNA tertiary structure; i.e. $\{i,j\} \in S$ if there is a hydrogen bond between the ith and jth nucleotide. The *page number* of S, denoted $\pi(S)$, is the minimum number k such that 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 two greedy algorithms, and show by an example that neither is optimal. We describe an algorithm running in time $O(n \log n)$ that produces a decomposition of an RNA structure S on n bases into at most $\omega(S) \cdot \log n$ disjoint secondary structures, where $\omega(S)$ denotes the maximum number of base pairs that may cross a given base pair. It follows that $\omega(S) \leq \pi(S) \leq \omega(S) \cdot \log n$, where $\pi(S)$ denotes the page number of S. We give an $O(n^{3/2})$ time algorithm for finding a 2-page decomposition of bisecondary structures for RNA sequences of size n, and we provide bounds on the *expected* page number of random structures having pseudoknots.

1 Introduction

Given an RNA sequence $\mathbf{s} = a_1, \ldots, a_n$, 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 $i < k < j < \ell$.

*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.

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 $i = k$.

For steric reasons, following convention, the threshold θ, or minimum number of unpaired bases in a hairpin loop, is taken to be 3. 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 S, then the ith and jth nucleotide can form a (possibly noncanonical) base pair. *(2)* If $\{i, j\}$ belongs to 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 [55], as implemented in `mfold` [54], `UNAFold` [27], `RNAfold` [18], `RNAstructure` [29], or *(ii) covariance model* approaches, such as the *stochastic context free grammar* approach implemented in `PFOLD` [22] and `tRNAscan-SE` [25]. 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 [30]; for a comparative benchmarking of a number of thermodynamics-based and covariance model methods, see the important study of Gardner et al. [11]. 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 [8]. This hybrid approach yields secondary structure accuracy of approximately 95%, comparable with the manually intensive method of *comparative sequence analysis* [16].

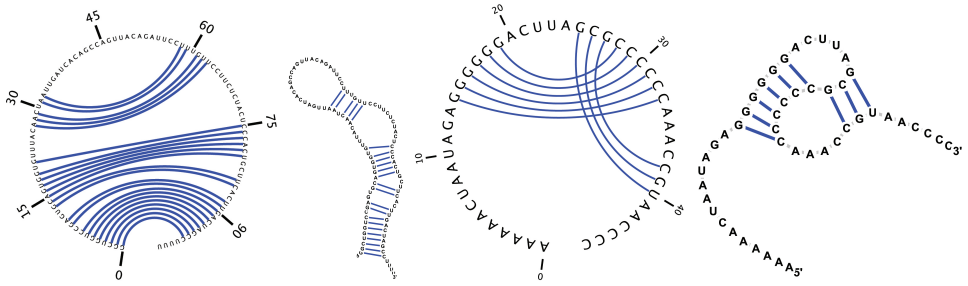

Figure 1: *(a,b)* Pseudoknot-free secondary structure of Y RNA with EMBL access code AAPY01489510/220-119, depicted in panel *(a)* in Feynman circular form, and in panel *(b)* in classical form. *(c,d)* Pseudoknotted structure for the Gag/pro ribosomal frameshift site of mouse mammary tumor virus, depicted in panel *(c)* in Feynman circular form, and in panel *(d)* in classical form. Images produced with sofware `jViz` [51] from structures taken respectively from Rfam [15] and Pseudobase [47].

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. Indeed, in the case that an RNA structure contains 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 [41] may yield predictions of different pseudoknotted regions! Another difficulty in pseudoknot structure prediction is the fact that there are no experi-

mentally determined free energies for pseudoknot formation, although Cao and Chen have described a computational method to approximate loop entropies for type-H pseudoknots [6]. Moreover, Lyngsø and Pedersen [26] have shown that minimum free energy pseudoknot structure prediction is an NP-complete problem. This situation is unfortunate, since pseudoknots often play important biological roles, such as promoting a programmed -1 ribosomal frameshift [45]. For additional biologically important examples of pseudoknot, consult the PSEUDOBASE database of pseudoknots [47, 44].

Pseudoknot prediction algorithms include the genetic algorithm of [1], the maximum weight matching approach of [43], the thermodynamics-based methods of [39, 9, 37, 38] which handle certain subclasses of pseudoknots, the Monte Carlo approaches of [32, 31], heuristics like position specific scoring matrices on tree structures [40] and `ProbKnot` [2], and the exact (exponential-time) methods using tree-width decomposition [53] and branch-and-bound [5]. Using tree-width decompositions, Huang et al. [19] developed fast and accurate genomic search for non-coding RNA pseudoknot structures.

1.1 Preliminaries and notation

A (general) RNA structure, defined in Section 1, can be identified with a simple, undirected graph G having vertices $1, 2, \ldots, n$ and undirected edges $\{i, j\}$, where $\max(i, j) > \min(i, j)-3$. An RNA structure thus uniquely corresponds to a given *contact map*, or *adjacency matrix*, $A = (a_{i,j})$, 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` [52] determines the collection of hydrogen bonds, including noncanonical bonds [24]. Thus for the purposes of this paper, an RNA structure is 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 $S \in \mathcal{S}(n)$, denoted by $\pi(S)$, is the minimum number n, such that S can be written as a disjoint union of n secondary structures; i.e. $S = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_n$, 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 (or adjacency graph) of a tertiary structure for the RNA sequence $\mathbf{s} = s_1, \ldots, s_n$; i.e. if there is a hydrogen bond between the ith and jth nucleotide, then $a_{i,j} = 1$, otherwise $a_{i,j} = 0$. Given RNA sequence $\mathbf{s}$ and contact map A, 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 $i \leq x \leq j \leq y$.

Let $\mathcal{P}_p(n)$ denote the set of RNA structures on n bases having page number at most p. A structure in $\mathcal{P}_p(n)$ can be visualized by writing the vertices $1, 2, \ldots, n$ along the spine of a book, where each of the p pages contains a (planar) secondary structure with no crossing edges. Figure 2 depicts a 3-page decomposition of an RNA structure, i.e. a structure belonging to $\mathcal{P}_3(n)$.

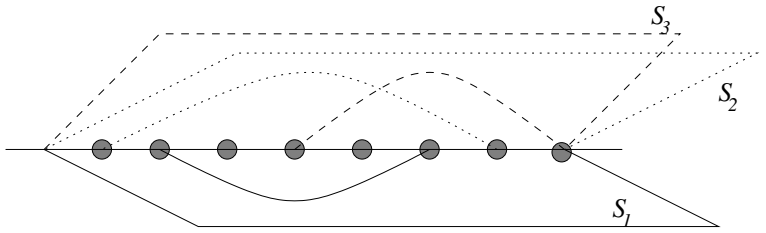

Figure 2: An RNA structure represented in three pages.

RNA structures are related to a class of graphs known in the literature as *circle graphs* [20].

Definition 2 (Circle graphs) *Consider a circle consisting of n points, arranged counterclockwise in order along the periphery of a circle. A circle graph $G = (V, E)$ consists of vertices $v \in V$, which are chords between these n points, and of edges $e \in E$ formed when two chords intersect.*

To each RNA structure S we associate its corresponding circle graph G_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 two 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 $i < j \leq i + 3$.

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 S, denoted by $\chi(S)$, 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 *$\chi(G_S)$ of the graph G_S.*

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

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

Clearly, $\omega(S)$ is the same as the size of the largest *clique* in G_S.

For each base pair b we can compute the number of base pairs crossing it, called the *crossing number* of b and denoted by $cn(b)$.

Definition 5 (Crossing number) *The crossing number of a general structure S, denoted by $cn(S)$, is the maximum of $cn(b)$ taken over all possible base pairs b in S.*

Figure 3 depicts a structure S such that $\omega(S) = 2 < \pi(S) = 3$. It is easy to construct

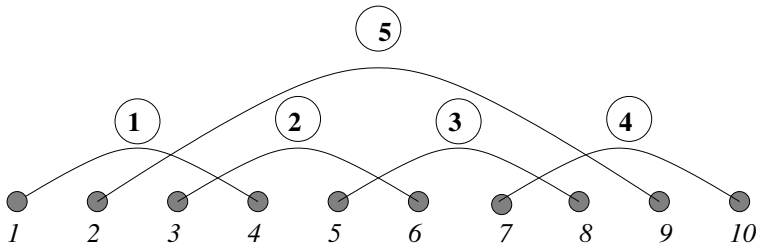

Figure 3: An RNA structure S on 10 bases with 5 base pairs, having clique number $\omega(S) = 2$ and page number $\pi(S) = 3$.

examples of structures with page number 2 and arbitrarily large crossing number.

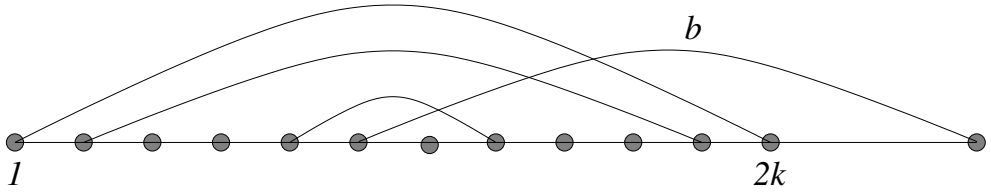

Figure 4: An RNA structure S on $2k$ bases with $k+1$ base pairs, having page number 2, yet the crossing number of base pair b is k.

It is straightforward to compute the clique number $\omega(S)$ of an RNA structure S. Simply scan vertices v one at a time, computing the clique number of all the base pairs $\{i, j\}$ such that $i < v < j$ and output the maximum number found. A similar observation holds for the crossing number $cn(S)$. However, no such simple method exists to compute the page number $\pi(S)$, since we show that page number is an NP-complete problem.

1.2 Related work

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. 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 [7] 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 [49, 50]. 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 [4]. Vernizzi et al. [49] developed recurrence relations to compute the number of genus g structures of a given RNA sequence, and Bon et al. [4] computed the genus of RNA X-ray structures, using the hydrogen bonding information provided by the program `RNAview` [52]. In his thesis, Bon [5] described a novel RNA energy model depending on topological genus, and using this energy model developed a branch-and-bound algorithm to compute the minimum energy pseudoknotted structure for a given RNA sequence.

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 [49] and the (general) treatment of genus for unordered graphs. In the latter case, Thomassen [46] has shown that computing the genus is NP-complete, although Filotti et al. [10] 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.

1.3 Outline of the paper

In this paper we address the question of computing the page number $\pi(S)$ of a given RNA structure S and producing a layout of the decomposition of S into $\pi(S)$ (planar) secondary structures. Section 2 shows that the problem of computing the page number is NP-complete for arbitrary structures with pseudoknots. Section 3 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. [4], where the topological genus is computed. Section 4 examines the performance of two greedy algorithms and provides an $\log n$ approximation algorithm for computing the page number. Section 5 considers the problem of finding a decomposition of a bisecondary structure. Section 6 gives bounds on the page and clique numbers of random RNA structures.

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 6 (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 S with pseudoknots on n bases; positive integer p.
Question: Is $\pi(S) \leq p$?

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 [12]; for a more recent and simpler proof of this fact see [28].

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. Let $G = (V, E)$ be a given circle graph, whose vertex set V consists of chords $\{i, j\}$ between distinct positions $1 \leq i < j \leq n$, ordered along the periphery of a circle, and whose edge set E consists of crossing chords $\{i, j\}$, $\{k, \ell\}$, where $i < k < j < \ell$. We describe how to transform G into a circle graph $G' = (V', E')$ of the same chromatic number, such that the vertex set V' consists of chords $\{i, j\}$ between distinct positions $1 \leq i < j \leq p(n)$, for a fixed polynomial p, where each position i belongs to at most one chord, thus permitting a labeling of positions in a manner that chords correspond to canonical base pairs. Additionally, we ensure that if chord $\{i, j\} \in V'$, then $|j-i| > \theta = 3$. In this fashion, we can ensure conditions (1,2,4) of the definition of secondary structure given in Section 1. Clearly then G' is the circle graph representation G_S for an RNA structure. 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.

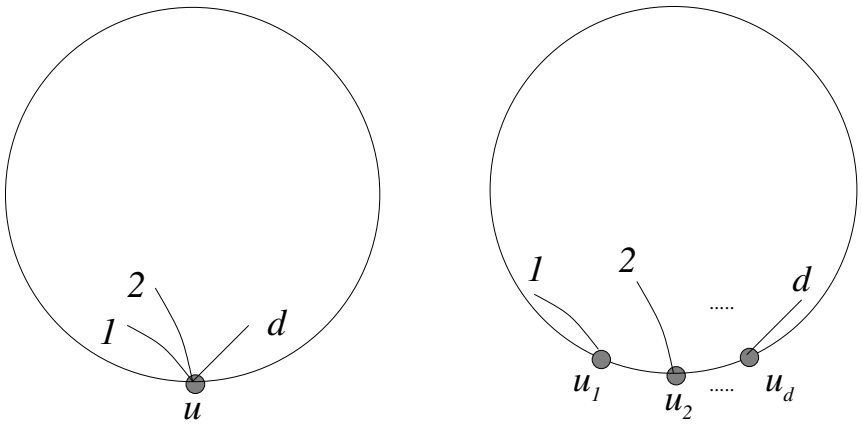

Figure 5: Polynomial time transformation of an arbitrary circle graph G (left panel) into a graph of the form G_S (right panel), for some pseudoknotted RNA structure, such that $\chi(G) = \chi(G_S)$.

We now describe the transformation for a given circle graph $G = (V, E)$, where $1, \ldots, n$ are positions occurring in counter-clockwise along the periphery of a circle, in which the vertices of G are chords. We define as follows a new circle graph $G' = (V', E')$ of the same chromatic number.

- **Positions:** For each position $1 \leq i \leq n$ along of the periphery of the circle used to define G, we associate positions $i_{j_1}, \ldots, i_{j_{\rho(i)}}$ in G', where $\{i, j_1\}, \ldots, \{i, j_{\rho(i)}\}$ is a listing

of all chords of G incident to i. If $1 \leq i \leq n$ is not incident to any chord of G, then we associate the position i_0 in G'. If $\{i, j\}$ is a chord of G where $i < j \leq i + \theta$, then we associate positions i_{-3}, i_{-2}, i_{-1} in G'. The positions of G' just defined are ordered lexicographically; i.e. $i_j < x_y$ if either $i < x$ or ($i = x$ and $j < y$).

- **Vertices:** For each vertex (chord) $\{i, j\}$ of G, we associate the vertex (chord) $\{i_j, j_i\}$ of G'.
- **Edges:** Edges of G' are defined by crossing chords.

See Figure 5 for an example of the transformation just defined. Clearly the transformation $G \rightarrow G'$ is computable in polynomial time, since the number of edges of G is at most $O(n^2)$. Moreover, G' is the representation of an RNA structure whose bases correspond to positions $1, \ldots, n'$ along the periphery of the circle corresponding to G'. Since each position is incident to at most one chord in G', the positions can be labeled by A,C,G,U in a manner that chords correspond to Watson-Crick or wobble pairs. If $\{i_j, j_i\}$ is a chord of G', then we have ensured that $|j_i - i_j| > \theta = 3$, hence the threshold requirement is met; moreover, by construction, there are no base triples. Finally, since ordering of nodes has been preserved, the chromatic number is as well, 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 [23] it is proved that $\chi(S) \leq 2^{\omega(S)}$. 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_S of a given RNA structure S has no triangles (clique of size 3; if there do not exist three distinct, mutually crossing base pairs in S.

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

Proof. If the RNA structure S has no triangles then [21] has shown that $\chi(S) \leq 8$. This was later improved to $\chi(S) \leq 5$ by Melnikov (see [20] for additional details). ■

3 Exact Computation of Page Number

Here we present an algorithm using *Constraint Programming* [48] to find the page number for a given RNA tertiary structure. Our approach is divided into three different steps:

1. Given a set S of (possibly noncanonical) base pairs $\{i, j\}$, we collapse 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.

3.1 Collapsing helices

Given a set S of (possibly noncanonical) base pairs $\{i, j\}$ from the tertiary structure of some RNA, as could be computed using the program `RNAview` [52], 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:

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

```
1.   collapseAndConstructGraph(S)
2.       S* = list of S in lexicographic order
3.       V = ∅
4.       E = ∅
5.       n = |S*|
6.       lastbp = S*[0]  first element of S*
7.       forall bp ∈ S* − {lastbp}
8.           if lastbp, bp are not consecutive
9.               V = V ∪ {lastbp}
10.              forall h ∈ V
11.                  if pseudoknot(h,bp)
12.                      E = E ∪ {h, h(bp)}
13.          lastbp = bp
14.      G = (V, E)
15.      return G
```

Figure 6: Algorithm to collapse base pairs and construct graph.

Figure 6 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 S.

3.2 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 [13] 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 [13]. 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 $D = \{0...|V| - 1\}$, and the domain for k to be $D(k) = \{1...|V|\}$.
- *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.

3.3 Results

In [4], 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 [4]. The results from Table 1 were obtained by our implementation of the above-described Constraint Programming algorithm in the COMET programming language [48].

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 [42] 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` [51], the 2-page pseudoknotted structure of 6TNA, computed by our program is displayed in Figure **??**. 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 **??**. 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.

4 Greedy and Approximation Algorithms for Computing the Page Number

In view of Theorem 1, computing the page number of an arbitrary RNA structure is NP-complete. In this section we provide greedy heuristics to compute for the decomposition of a given RNA structure into a possibly suboptimal number of pages. Additionally, we provide a $\log n$ approximation algorithm to compute the page number of an RNA structure.

4.1 Greedy algorithm

An obvious greedy algorithm to compute an upper bound for page number proceeds as follows: *(1)* order the base pairs of the RNA structure, and *(2)* place the base pairs on pages in this order by adding new pages as necessary to prevent crossings. Suppose we are given an RNA structure S with bases $1, 2, \ldots, n$. Partition as follows the collection of base pairs of S into disjoint secondary structures $S_1, S_2, \ldots, S_k$.

4.1.1 Algorithm *GPN* (Greedy Page Number)

Enumerate all base pairs of S as $b_1, b_2, \ldots, b_m$ and place b_1 into the first secondary structure 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 $S_1, S_2, \ldots, S_j$. Given the next base pair b_{i+1}

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. [4] computed the topological genus. In each case, the exact page number was computed by our Constraint Programming algorithm within one second.

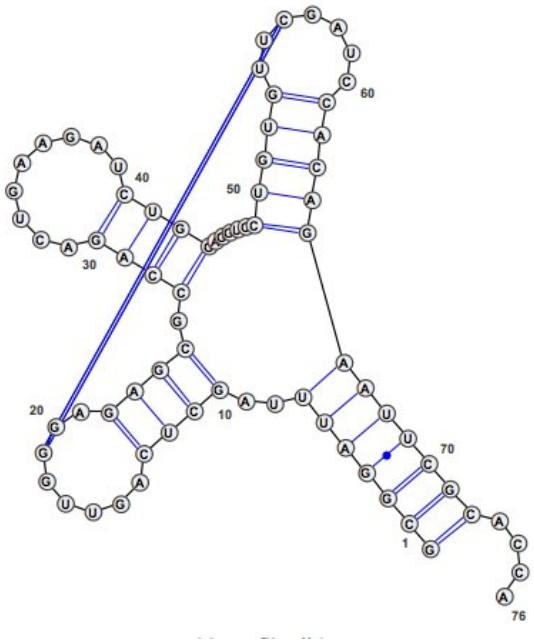

Figure 7: Pseudoknotted structure of the 76 nt yeast phenylalanine transfer RNA with PDB code 6TNA [42]. This example is trivial, since there is only one pseudoknot. Image produced using `jViz` [51].

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

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

It is clear that the secondary structures $S_1, S_2, \ldots, S_k$ constructed have no crossings and $\pi(S) \leq k$, thus the following result is immediate.

Theorem 3 *Algorithm GPN is a greedy algorithm which when given an RNA structure S computes a decomposition into secondary structures. The running time of the algorithm is $O(n^2)$, where n is the number of bases of the structure S.* ■

4.1.2 Greedy is not optimal

In the sequel we give two examples indicating the failure of the greedy algorithm to construct an optimal page decomposition of a given RNA structure.

Example 1 The following example shows that the greedy algorithm GPN is not optimal in the sense that for a given $S \in \mathcal{S}(n)$ the resulting number of secondary structures can be greater than the optimal $\pi(S)$. To see this take ten bases numbered $1, 2, \ldots, 10$. Consider the following five base pairs

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

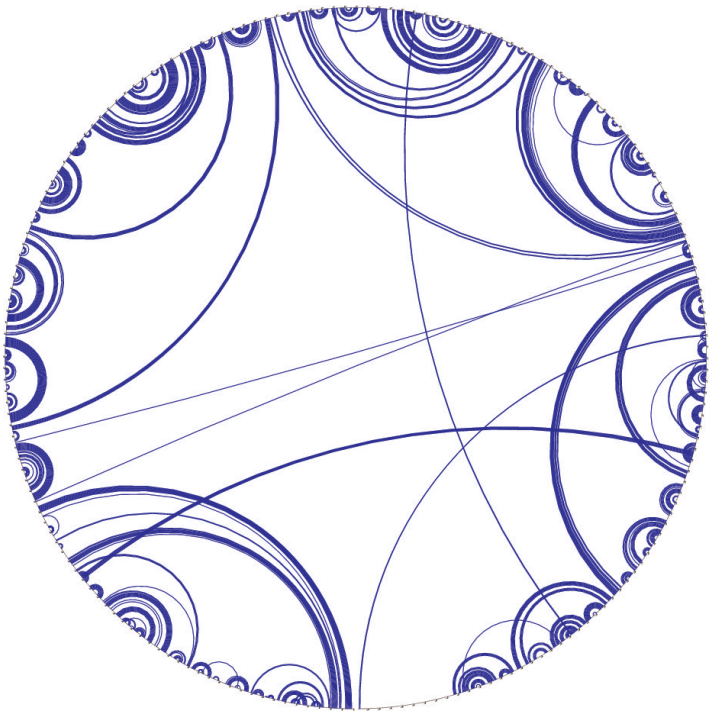

Figure 8: Pseudoknotted structure of the 2922 nt sequence of 23S ribosomal RNA with PDB code 1KC8:A (A chain of 1KC8). Minimum page number layout produced by the Constraint Programming method described in this section; image produced using `jViz` [51].

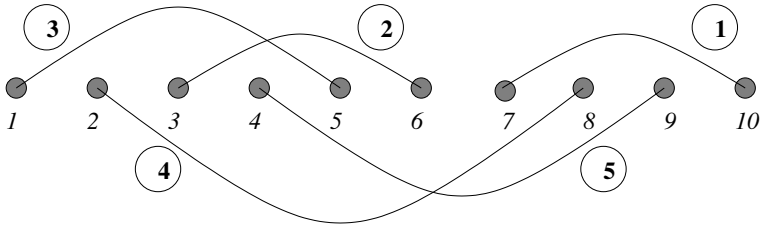

Figure 9: Example of an RNA structure with ten bases and page number three, but for which the greedy algorithm allocates four pages.

and suppose that the base pairs are being considered in this order. The greedy algorithm will output $S_1 = \{\{7, 10\}, \{3, 6\}\}$. Since $\{1, 5\}$ crosses one of the base pairs of S_1 we must have that $S_2 = \{\{1, 5\}\}$. Since $\{2, 8\}$ crosses base pairs in both S_1 and S_2 we have that $S_3 = \{\{2, 8\}\}$. Finally, since $\{4, 9\}$ crosses base pairs in S_1, S_2 and S_3 we must have that $S_4 = \{\{4, 9\}\}$. Therefore the greedy algorithm allocates four pages. Nevertheless three pages are sufficient, namely

$$S_1 = \{\{7, 10\}, \{1, 5\}\}, S_2 = \{\{3, 6\}, \{2, 8\}\}, S_3 = \{\{4, 9\}\},$$

and the page number of this RNA structure is three.

In [36], Ponty described a simple modification of the Nussinov-Jacobson algorithm [35],

to compute the maximum size secondary structure (collection of nested base pairs) contained within a given RNA structure. Independently and later, Smit et al. [41] rediscovered the same algorithm to compute the maximum planar portion of an RNA structure. One could consider a a second type of greedy algorithm, obtained by iteratively applying Ponty's maximum planarization to a given RNA structure, where each successive maximum planarization appears on a separate page. The previously given example also shows that this second type of greedy algorithm may output a larger number of pages than the optimal. Indeed, the maximum planarization greedy approach would place base pairs $1, 2$ on page 1, base pair 5 on page 2, base pair 4 on page 3, and base pair 3 on page 4.

Example 2 The following example shows that the greedy algorithm GPN may assign $O(\log n)$ for an RNA structure which requires only a constant number of pages. The sets of base pairs are constructed recursively in stages $S_0, S_1, \ldots, S_n$. Initially, S_0 consists of one base pair. Assume that S_n has been constructed. Then the set S_{n+1} is constructed by appending a copy of S_n to itself plus the addition of a base pair which 1) surrounds the second copy of S_n, 2) crosses the first copy of S_n, 3) its leftmost base is enclosed inside the rightmost innermost base pair of the first copy of S_n. The resulting structure is depicted in Figure 11 for $n = 0, 1, 2$.

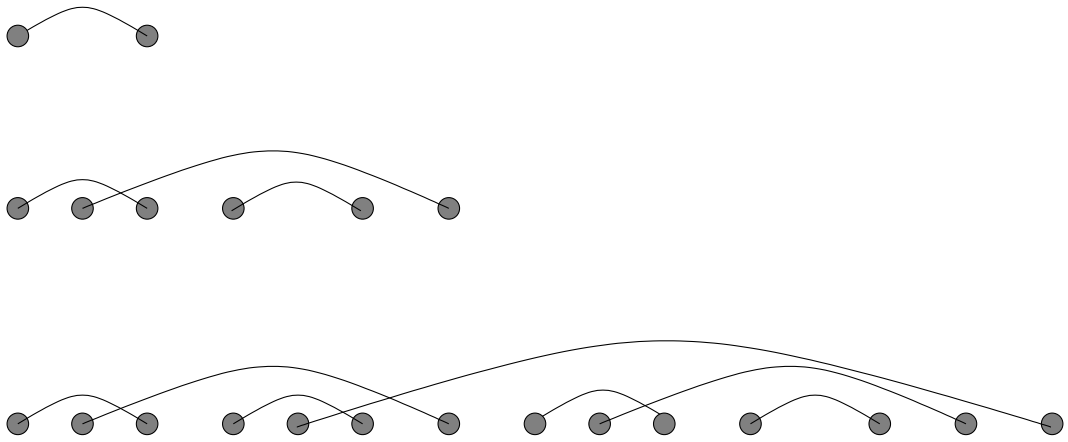

Figure 10: Example of the sequence of sets of base pairs S_0, S_1, S_2.

The number B_n of base pairs introduced by the n-th 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 GPN. Enumerate the base pairs by occurrence of their leftmost base, say $b_0, b_1, b_2, \ldots$. At the k-th 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 GPN.

4.2 Approximation algorithm

In this section, we describe a method that provides a upper bound for page number within a factor of $\log n$.

Theorem 4 *There is a* $\log n$ *approximation algorithm producing a decomposition of an RNA structure* S *on* n *bases into* $O(\pi(S)\log n)$ *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 $\omega(S)$ pages, where $\omega(S)$ was previously defined as the clique number of S, and therefore $\pi(S) \geq \omega(S)$. Next we give a page decomposition algorithm and prove that

$$\pi(S) \leq \omega(S) \cdot \log n.$$

This and the fact that $\pi(S) \geq \omega(S)$ proves the approximation claim for $\pi(S)$ in the theorem.

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 11, where for simplicity, we assume that n is even.

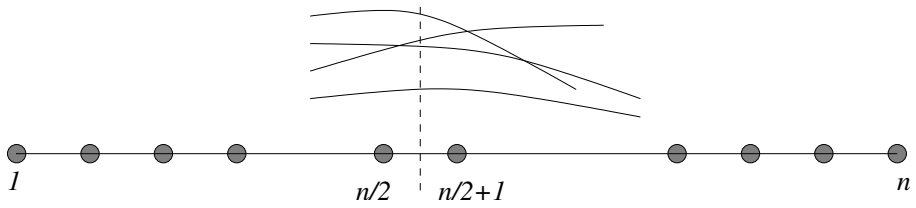

Figure 11: Using divide and conquer in a structure with pseudoknots.

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 $\omega(S)$, thus resulting in at most $\omega(S)$ pages. This coloring can be found in time $O(n\log n)$. Indeed, the graph induced on the set of edges connecting bases between $1..\lfloor n/2 \rfloor$ and $\lfloor n/2 \rfloor + 1..n$ has the property that each of its edges is crossing the vertical dashed line depicted in Figure 11 and as such is a permutation graph (see [14]). Now remove all the base pairs crossing this vertical line. Two RNA structures result, the first from $1..\lfloor n/2 \rfloor$, and the second from $\lfloor n/2 \rfloor + 1..n$. Since the base pairs of $1..\lfloor n/2 \rfloor$ and $\lfloor n/2 \rfloor + 1..n$ do not interfere with each other, the original structure can be colored with the $\omega(S)$ 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 $1..n/2$ and $n/2+1..n$. If $\chi(n)$ is the minimum number of colors required to color an RNA structure on n bases then it follows that

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

Applying this technique recursively we derive that $\pi(S) \leq \omega(S)\log n$, as desired. This proves the approximation claim in the theorem.

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 $1..\lfloor n/2 \rfloor$ and $\lfloor n/2 \rfloor + 1..n$. Algorithms for finding such a clique are known and require time $O(n\log n)$ (see [14] for more details on algorithms for permutation graphs). ■

5 Bisecondary structures

An RNA structure is a *bisecondary structure*, first defined in [17], if it is the disjoint union of at most two secondary structures; i.e. if it has page number at most 2. In this section we

limit ourselves to bisecondary structures.

5.1 Finding a decomposition

An interesting question arises as to how to compute the decomposition of a structure S, given that we know the page number $\pi(S)$. We address this question below by providing an efficient algorithm when the page number is at most 2.

Theorem 5 *Assume that S is a structure on n bases such that $\pi(S) \leq 2$. There is an algorithm with running time $O(n^{3/2})$ to compute the decomposition of S into the minimal number of pages.*

Proof. Consider the circle graph G_S associated with the structure S. It is clear that G_S is bipartite. The partition of the vertex set of G_S into two parts is according to the decomposition of S into the two pages. Now observe that finding a decomposition is the same as finding a matching in the bipartite graph: since all cycles are even we can partition the vertex set of the graph G_S according to the parity of a vertex in a cycle. The complexity bound $O(n^{3/2})$ follows immediately from [33]. This completes the proof of the theorem. ■

6 Random structures with pseudoknots

Next we consider the clique size and page number for random RNA structures of a certain type. We begin by defining the particular notion of *random structure* S considered in this section, where S has exactly n base pairs and at most $2n$ nucleotides incident to the base pairs (base triples, etc. are allowed). Suppose we have n points on the periphery of a circle. Assume that n chords are drawn randomly and independently, each with probability $\frac{1}{\binom{n}{2}}$. For a fixed node u, there are $2(n-1)$ *ordered* base pairs (i,j) incident to u, i.e. in which $u \in \{i,j\}$, compared with a total of $n(n-1)$ possible ordered base pairs. It follows that the probability, for fixed node u, of selecting an *ordered* base pair (i,j) incident to u is $\frac{2(n-1)}{\binom{n}{2}} = \frac{2}{n}$; hence the probability of selecting an *unordered* base pair incident to u is $\frac{1}{n}$, as two ordered base pairs correspond to each unordered base pair. Hence, the probability distribution $p_k := \Pr[\text{degree of point } u \text{ is } k]$ of the *degree* of a given point is Bernoulli, i.e.,

$$p_k = \binom{n}{k}(1/n)^k(1-1/n)^{n-k},$$

which gives expected degree equal to $\sum_{k=0}^{n} kp_k = n(1/n) = 1$. We now construct as follows a random RNA structure S with n base pairs and no base triples. For each vertex which is the endpoint of t chords, $t > 1$, we add $t-1$ vertices and associate the chords with each of the vertices, one-by-one, in such a way that chords do not overlap. Additionally, we ensure the threshold condition, that if $\{i,j\}$ is a base pair, then $|j-i| > \theta = 3$ – see details from construction in Theorem 1. Clearly, the resulting RNA structure has the same number of base pairs and at most $2n$ bases; moreover, the clique number of S remains unchanged. In the sequel we determine the clique size of this random RNA structure.

Theorem 6 *For the random RNA structure S_n defined above we can prove that*

1. $E[\omega(S_n)] \in \Omega(\sqrt{n/\log n})$,

2. $E[\omega(S_n)] \in O(\sqrt{n})$, *and as a consequence*

3. $E[\pi(S_n)] \in \Omega(\sqrt{n/\log n})$ *and* $E[\pi(S_n)] \in O(\sqrt{n}\log n)$.

Proof. Consider a random RNA structure $S := S_n$ defined above. The lower bound in Part 3 of the theorem follows from the fact that $\omega(S) \leq \pi(S)$, while the upper bound follows directly from Theorem 4 since $\pi(S) \leq \omega(S) \cdot \log n$. It remains to prove Parts 1, 2 of the theorem.
Proof of Part 1. Without loss of generality assume that k is an integer such that $2k$ divides n. Divide up the circle into $2k$ consecutive arcs each of size $\frac{n}{2k}$ (see Figure 12) and number

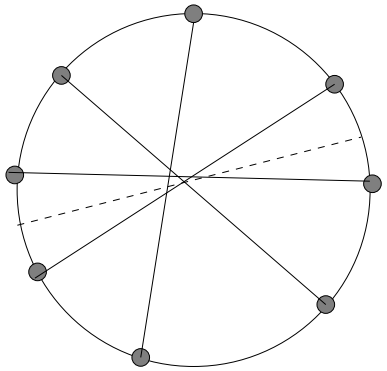

Figure 12: Depicted is a clique consisting of 4 chords (solid lines). A new chord (dashed line) increments the size of the clique to 5 if and only if its vertices lie in an antipodal pair of arcs.

these arcs counter-clockwise $S_1, S_2, \ldots, S_{2k}$. We can view the "antipodal" pair $\{S_i, S_{i+k}\}$ as a "bucket". Thus we have k buckets and n chords. It is clear that if every bucket has a chord in it then by selecting one chord from each bucket the resulting set of chords forms a clique. These chords are thrown randomly and independently into the buckets and we would like to guarantee with high probability that every bucket has a chord. We can guarantee this with high probability using the probabilities associated with the coupon collector's problem, provided that a total of at least $k \log k$ chords fall into the buckets (see [34]). Clearly, the probability that a random chord falls into a given bucket is equal to

$$\frac{\left(\frac{n}{2k}\right)^2}{\binom{n}{2}} \approx \frac{1}{2k^2}.$$

As a consequence, the probability that a random chord falls into at least one of the k buckets is at most $\sum_{i=1}^{k} \frac{1}{2k^2} = k\frac{1}{2k^2} = \frac{1}{2k}$.

In the experiment above we have n chords (drawn independently) and k buckets. It is therefore clear that with high probability, an expected fraction $\frac{n}{k}$ of these chords will fall into the k buckets, while the remaining $n - n/k = n(1 - 1/k)$ will fall outside these k buckets. Moreover, the previous argument shows that if

$$\frac{n}{k} > k \log k \tag{1}$$

then with high probability every bucket has a chord and consequently there is a clique of size k. Since $k = \sqrt{n/\log n}$ satisfies Inequality 1 we conclude the desired lower bound claimed in Part 1.

Proof of Part 2. To prove the upper bound $O(\sqrt{n})$ stated in Part 2 we argue as follows. First we show by induction on k that the probability that a set of k chords forms a clique is at most $1/(k-1)!$. Indeed, consider the event E_{k+1} that a random set $\{e_1, \ldots, e_k, e_{k+1}\}$ of $k+1$ chords forms a clique. Also consider the event F_k that e_{k+1} crosses all the chords $e_1, \ldots, e_k$. Now observe that

$$\begin{aligned}
\Pr[E_{k+1}] &= \Pr[E_k \ \& \ F_k] \\
&= \Pr[F_k \mid E_k] \cdot \Pr[E_k] \\
&= \Pr[F_k \mid E_k] \cdot \frac{1}{(k-1)!} \text{ (By induction Hypothesis)} \\
&= \left(\frac{1}{\binom{n}{2}} \sum_{i=1}^{k} t_i t_{i+k} \right) \cdot \frac{1}{(k-1)!}.
\end{aligned}$$

The last identity is valid because the k chords forming a clique divide up the circle into $2k$ subintervals of respective, successive lengths $t_1, t_2, \ldots, t_{2k}$ (see Figure 12) such that $\sum_{i=1}^{2k} t_i = n - 2k$ and the additional chord e_{k+1} forms a clique together with the chords $e_1, e_2, \ldots, e_k$ if its endpoints are chosen to an antipodal pair. Therefore the sum in the right-hand side is maximixed when $t_i = \frac{n}{2k}$ in which case

$$\Pr[E_{k+1}] \leq \left(\frac{1}{\binom{n}{2}} \sum_{i=1}^{k} \left(\frac{n}{2k} \right)^2 \right) \cdot \frac{1}{(k-1)!} \leq \frac{1}{k!},$$

which completes the inductive proof. It follows that

$$\begin{aligned}
\Pr[\exists(\text{a clique of size } k)] &\leq \binom{n}{k} \Pr[C \text{ is a clique of size } k] \\
&\leq \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot \frac{1}{k!} \\
&\leq \frac{n^k}{(k!)^2} \\
&\leq \frac{n^k}{(k/e)^{2k} 2\pi k} \text{ (Using Stirling's Formula)} \\
&\leq \left(\frac{n}{k^2} \right)^k \frac{e^{2k}}{2\pi k} \\
&\leq \frac{1}{e^k 2\pi k},
\end{aligned}$$

provided that $k \geq e^{3/2}\sqrt{n}$. It follows from this that

$$\begin{aligned}
\Pr[\exists(\text{a clique of size} \geq e^{3/2}\sqrt{n})] &\leq \sum_{k \geq e^{3/2}\sqrt{n}}^{n} \Pr[\exists(\text{a clique of size } k)] \\
&\leq n e^{-e^{3/2}\sqrt{n}}.
\end{aligned}$$

Finally, concerning the upper bound on the expected value, we observe that

$$\begin{aligned} E[\omega(S)] &= \sum_k \Pr[\omega(S) \geq k] \\ &= \sum_{k \leq e^{3/2}\sqrt{n}} \Pr[\omega(S) \geq k] + \sum_{k \geq e^{3/2}\sqrt{n}} \Pr[\omega(S) \geq k] \\ &\leq e^{3/2}\sqrt{n} + n^2 e^{-e^{3/2}\sqrt{n}} \\ &\in O(\sqrt{n}). \end{aligned}$$

This completes the proof of Theorem 6. ■

7 Conclusion

In this paper, we have proven that that computing the page number of possibly pseudoknotted RNA structures is NP-complete. We described two greedy algorithms (GMP and an iteration of Ponty's maximum planarization), and showed by an example that neither is optimal. (Of course, the non-optimality of any polynomial time algorithm, in particular of the greedy algorithm, follows immediately from NP-completeness of the page number.) We have described an $O(n^{3/2})$ time algorithm to determine a $\leq$ 2-page decomposition of bisecondary structures for RNA sequences of size n, and we have provided bounds on the *expected* page number of random RNA structures constructed by randomly choosing n base pairs.

The application of topological genus was introduced in the context of RNA structure in [49, 50], and subsequently used in [4] to classify tertiary RNA structures from the Protein Database [3]. In [5], an energy model for RNA structures was introduced which includes a pseuodoknot penalty according to genus. Since genus is a topological notion, that does not take into account stericity and other molecular constraints, one might consider whether page number provides a better classification of RNA structures. Unfortunately, the work in this paper shows that no reasonable energy model involving page number exists, since page number is an NP-complete problem.

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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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.

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

[19] 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.

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

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

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

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

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

[25] 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.

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

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

[28] D. Marx. A short proof of the NP-completeness of circular arc coloring, 2003. 7th November (unpublished), `http://www.cs.bme.hu/~dmarx/papers/circularNP.pdf`.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1998.

[35] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proceedings of the National Academy of Sciences, USA*, 77(11):6309–6313, 1980.

[36] Y. Ponty. Modélisation de séquences génomiques structurées,génération aléatoire et applications, 2006.

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

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

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

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

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

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.