Linear Time Algorithm for Parsing RNA Secondary Structure

Extended Abstract

Baharak Rastegari and Anne Condon

Department of Computer Science, University of British Columbia

Abstract. Accurate prediction of pseudoknotted RNA secondary structure is an important computational challenge. Typical prediction algorithms aim to find a structure with minimum free energy according to some thermodynamic ("sum of loop energies") model that is implicit in the recurrences of the algorithm. However, a clear definition of what exactly are the loops and stems in pseudoknotted structures, and their associated energies, has been lacking.

We present a comprehensive classification of loops in pseudoknotted RNA secondary structures. Building on an algorithm of Bader et al. [2] we obtain a linear time algorithm for parsing a secondary structures into its component loops.

We also give a linear time algorithm to calculate the free energy of a pseudoknotted secondary structure. This is useful for heuristic prediction algorithms which are widely used since (pseudoknotted) RNA secondary structure prediction is NP-hard. Finally, we give a linear time algorithm to test whether a secondary structure is in the class handled by Akutsu's algorithm [1]. Using our tests, we analyze the generality of Akutsu's algorithm for real biological structures.

1 Introduction

RNA molecules play diverse roles in the cell: as carriers of information, catalysts in cellular processes, and mediators in determining the expression level of genes [8]. The structure of an RNA molecule is often the key to its function with other molecules. In particular, the *secondary structure*, which describes which bases of an RNA molecule bond with each other, can provide much useful insight as to the function of the molecule. If the RNA molecule is viewed as an ordered sequence of n bases (Adenine (A), Guanine (G), Cytosine (C), and Uracil (U)), indexed starting at 1 from the so-called 5' end of the molecule, then its secondary structure is a set of pairs $i \cdot j$, $1 \leq i < j \leq n$ with each index in at most one pair.

Most well known are *pseudoknot free* secondary structures in which no base pairs overlap - that is, there do not exist two base pairs $i \cdot j$ and $i' \cdot j'$ in the structure with $i < i' < j < j'$. Because of their biological importance, there has been a huge investment in understanding the thermodynamics of pseudoknot free secondary structure formation. For example, it is well understood that in a

pseudoknot free secondary structure, the base pairs together with unpaired bases form hairpin loops, internal loops (of which stacked pairs and bulge loops are special cases), external loops, or multiloops, with every unpaired base in exactly one loop and every base pair in exactly two loops. Parameters for estimating the free energies of such loops have been determined experimentally. The standard thermodynamic model posits that the free energy of a pseudoknot free secondary structure is the sum of the energies of its loops. A pseudoknot free secondary structure can be conveniently represented as a string in dot-parenthesis format, a generalization of a string of balanced parentheses in which matching parentheses denote base pairs and dots denote unpaired bases. It is straightforward to parse a pseudoknot free secondary structure represented in dot-parenthesis notation in linear time, in order to determine its loops and calculate its free energy. Finally, dynamic programming algorithms can find the minimum free energy (mfe) pseudoknot free secondary structure in $O(n^3)$ time; the mfe structure is the most stable of the possibly exponentially many structures that a molecule may form, according to current models.

In contrast, there has been no classification of loops in pseudoknotted secondary structures, though some examples of structural motifs, such as kissing hairpins, have been named. Since pseudoknotted secondary structure prediction is NP-hard, several polynomial time algorithms have been proposed for predicting the mfe secondary structure from restricted classes of structures that may contain pseudoknots. Of these, the $O(n^6)$ algorithm of Rivas and Eddy [12] handles (i.e. finds the mfe structure from) the most general class of structures. However, the loop types and thermodynamic model underlying the Rivas and Eddy and other algorithms are specified only implicitly in the recurrence equations of the algorithms. There is not a one-to-one correspondence between loops and terms in the recurrence equations, making it difficult to infer the loop types directly from the recurrences. The underlying energy models are unclear; there has been no algorithm to calculate the energy of a structure, and no way to compare the quality of thermodynamic models proposed by different authors.

In this work we present the first classification of loops that arise in pseudoknotted secondary structures. Our classification is derived from the algorithm of Rivas and Eddy, and allows us to formulate the thermodynamic models underlying the Rivas and Eddy and other dynamic programming algorithms as sum-of-loop-energies models. With this description, it becomes possible to evaluate the strengths and weaknesses of current thermodynamic models for pseudoknotted structures.

By extending an algorithm of Bader et al. [2], it is possible to parse a given secondary structure into its component loops in linear time. We present two applications of this parsing algorithm. First, we show how to calculate the free energy of a pseudoknotted secondary structure in linear time. This can be useful in heuristic algorithms, which hold promise since pseudoknotted secondary structure prediction is NP-hard [11].

The second application of our parsing algorithm is in assessing the trade-off between generality and running time of dynamic programming algorithms for

RNA secondary structure prediction. Each dynamic programming algorithm in the literature only predicts structures from a restricted class. Usually, the more general the class, the higher the running time of the algorithm. An outstanding challenge is to design efficient dynamic programming algorithms that can predict biologically important structures. For example, Akutsu [1] proposed an algorithm that runs in $O(n^5)$ time, can in theory handle more secondary structures than the $O(n^5)$ algorithm of Dirks and Pierce [9], though less than the $O(n^6)$ algorithm of Rivas and Eddy. As another example, Uemura [14] proposed an algorithm that runs in $O(n^5)$ time, similar to Akutsu's algorithm in time complexity, but in theory handle more secondary structures than Akutsu's algorithm, though it is much more harder to understand and analyse. Let U, A, D&P, and R&E denote the classes of structures handled by the Uemura, Akutsu, Dirks and Pierce, and Rivas and Eddy algorithms, respectively. The question we address is: does A contain more biologically meaningful structures than does D&P and perhaps as many as U and/or R&E?

To help answer this question, we apply the parsing algorithm to give linear time test for membership in class A. In previous work [7], we obtained linear time tests for membership in the D&P and R&E classes. We provide a comparison of all four algorithms on a set of 1439 biological structures; the result shows that exactly 2 of the structures are in class A but not in class D&P.

The paper is organized as follows. In Sec. 2, we define what is a closed region in an RNA secondary structure. (The parsing algorithm, based on an algorithm of Bader et al. [2], is not shown due to the lack of space). In Sect. 3 we present our loops classification and our algorithm for enumerating the loops of a secondary structure. We briefly describe how to calculate the free energy of a secondary structure in Sect. 4. Our algorithm for testing membership in Akutsu's class is in Sect. 5, and conclusions are in Sect. 6.

We should note that some details of the algorithms and most of the details of the proofs are eliminated in this extended abstract.

2 Closed Regions

Here we first introduce *closed regions* of a secondary structure, which are important throughout the paper. Examples are shown in Fig. 1, where a secondary structure is represented as an arc diagram, in which base indices are shown as vertices on a straight line (backbone), ordered from the $5'$ end, and arcs (always above the straight line) indicate base pairs. Intuitively, a closed region is a "minimal" set of contiguous base indices - corresponding to a region of the line - with the property that no arcs leave the region and there is at least one arc in the region. The definitions in this and the following sections are with respect to a fixed non-empty secondary structure R for an RNA sequence of length n.

We denote the set of indices $i, i+1, ..., j$ by $[i; j]$ and call this set a region if $i \leq j$. We say that region $[i; j]$ is **weakly closed** if it contains at least one base pair and for all base pairs $i' \cdot j'$ of R, $i' \in [i; j]$ if and only if $j' \in [i; j]$. We say that $[i; j]$ is **closed**, and write $i; j$, if either (i) $i = 1$ and $j = n$ or (ii) $[i; j]$ is

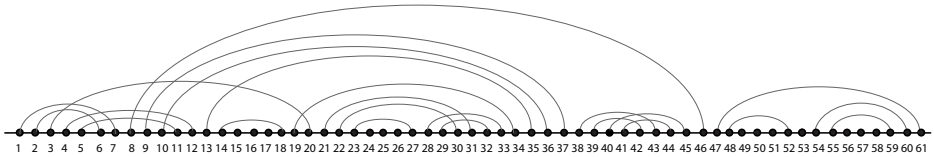

Fig. 1. Arc diagram representation of an RNA secondary structure R. **(a)** $[1;46]$, $[38;45]$, and $[47;61]$ are closed regions. $[48;60]$ is weakly closed but it is not closed as $[48;52]$ is weakly closed. $[38;45]$ is a pseudoknotted closed region. 38.43 and 40.45 are its external base pairs, and 38 and 45 are its left and right borders respectively. 38.43, 39.42, 40.45 and 41.44 are all pseudoknotted pairs and $38, 39, ..., 45$ are all pseudoknotted bases. **(b)** $[48;52]$ and $[54;60]$ are disjoint closed regions and both are nested in $[47;61]$. **(c)** $[8;10] \cup [36;46]$ is a band of pseudoknotted closed region $[1;46]$, and 8.46 and 10.36 are the band's outer and inner closing pairs. $[8;10]$ and $[36;46]$ are the band's regions, and 8 and 46 are the left and the right border of the band. 8.46, 9.37 and 10.36 span the band. **(d)** 47.61 is a multiloop external base pair with $(48, 52)$ and $(54, 60)$ as tuples. $(1, 46)$ and $(47, 61)$ are the tuples of an external loop. **(e)** $[1;46]$ is a pseudoknotted loop with bands $[1;2] \cup [6;7]$, $[3;3] \cup [20;20]$, $[4;5] \cup [11;12]$, $[8;10] \cup [36;46]$ and $[13;19] \cup [34;35]$. $[21;33]$ is the closed region nested in $[1;46]$. **(f)** 1.7 and 2.6 are the external and internal base pairs of an interior-pseudoknotted loop. **(g)** 8.46 is the external base pair of a multi-pseudoknotted loop with $(9, 37)$ and $(38, 45)$ as tuples. **(h)** $[38;45]$ is an in-Band loop, $[21;33]$ is an out-Band loop, and 8.46 is the external base pair of a span-Band loop (multi-pseudoknotted loop).

weakly closed and for all l with $i < l < j$, $[i;l]$ and $[l;j]$ are not weakly closed (Fig. 1(a)).

Let $i;j'$. If i' and j are such that $i.j$ and $i'.j'$ then we say that $i.j$ and $i'.j'$ are the *external* base pairs of $[i;j']$. If $i.j'$ then the region has just one external base pair; otherwise we call $[i;j]$ a **pseudoknotted closed region**. We also refer to i and j' as $[i;j']$'s left and right borders respectively.

Pair $i.j$ is **pseudoknotted** if there exists $i'.j'$ with $i < i' < j < j'$ or $i' < i < j' < j$. We also refer to i and j as **pseudoknotted** base indices.

2.1 Closed Regions Tree

Let $i;j$ and $i';j'$ with $i < i'$. If $j < i'$ we say that $[i;j]$ and $[i';j']$ are *disjoint*; otherwise we say that $[i';j']$ is *nested* in $[i;j]$ (Fig. 1(b)).

We say that closed region $[i';j']$ is a child of closed region $[i;j]$ if $[i';j']$ is nested in $[i;j]$ and is not nested in any closed region $[i'';j'']$ with $i < i''$. We say that $[i;j]$ and $[i';j']$ are siblings if they are children of the same closed region and $i \neq i'$. So the closed regions form a tree structure.

A tree $T(R)$ in which the children of a node are ordered is called the closed regions tree of R if: (i) there is a 1-1 correspondence between nodes of the tree and closed regions of R, and (ii) if node V corresponds to closed region C then V is the parent of all the nodes whose corresponding closed regions are nested in C. The children of each node are ordered by the left index of the closed region.

Building on an algorithm by Bader et al. [2], parsing algorithm (not shown) builds the closed region tree in linear time. (Details will be in full paper - omitted in this extended abstract)

3 Loops

In this section we describe the loops that comprise a pseudoknotted secondary structure, and how these can be enumerated in linear time. Models underlying the algorithm of Rivas and Eddy [12] and the algorithm of Dirks and Pierce [9] can be expressed by sum of the loops that we describe here. We need one important definition, that of a *band*.

3.1 Bands

Loosely speaking, a band is a pseudoknotted stem, which may contain internal loops or multi loops (Fig. 1(c)). We next define a band formally.

Let $i_2.j_2$ be a pseudoknotted base pair. We say that $i_2.j_2$ is *directly* banded in $i_1.j_1$ if (i) $i_1 \leq i_2 < j_2 \leq j_1$, and (ii) $[i_1+1, i_2-1]$ and $[j_2+1, j_1-1]$ are weakly closed. Note that the "is directly banded in" relation is reflexive. We let "are banded" be the symmetric and transitive closure of the "is directly banded in" relation. Let B be an equivalence class under the "are banded" relation. That is, B is a set of base pairs such that every two base pairs in B are banded and every base pair in B is pseudoknotted. B has outer and inner closing base pairs $i_1 \cdot j_1$ and $i'_1 \cdot j'_1$ respectively, such that for every base pair $i.j$ in B, $i_1 \leq i \leq i'_1$ and $j'_1 \leq j \leq j_1$. Note that $i_1 \cdot j_1$ may equal $i'_1 \cdot j'_1$.

We call the union of two non-overlapping regions a *gapped region*. A gapped region $[i_1; i'_1] \cup [j'_1; j_1]$ is a *band* if for some equivalence class B, $i_1 \cdot j_1$ and $i'_1 \cdot j'_1$ are the closing pairs of B. We refer to i_1 and j_1 as the left and the right border of the band respectively (Fig. 1(c)).

We refer to $[i_1; i'_1]$ and $[j'_1; j_1]$ as the band regions, which have borders i_1, i'_1 and j'_1, j_1 respectively. Closed region $i; j$ is *contained in* band $[i_1; i'_1] \cup [j'_1; j_1]$, if and only if $i; j$ is in a band region - that is, $i, j \in [i_1; i'_1]$ or $i, j \in [j'_1; j_1]$ - and there is no p, q with $p; q$, $p < i < j < q$, such that $p, q \in [i_1; i'_1]$ or $p, q \in [j'_1; j_1]$. Base pair $i.j$ *spans* band $[i_1; i'_1] \cup [j'_1; j_1]$ if $i_1 \leq i \leq i'_1$ and $j'_1 \leq j \leq j_1$.

We say that $[i_1; i'_1] \cup [j'_1; j_1]$ is a *band of closed region* $[i; j]$ if $i \leq i_1 \leq j_1 \leq j$ and there is no $p; q$ with $i < p \leq i_1 < j_1 \leq q < j$.

Lemma 1. *Let $i_1.j_1, i_2.j_2, ..., i_n.j_n$, $i_1 < i_2 < ... < i_n$, be the base pairs that span band $[i_1; i'_1] \cup [j'_1; j_1]$. Then $j_n < .... < j_2 < j_1$.*

3.2 Loop Types

Our definitions of hairpin and interior loops are standard for pseudoknot free structures so we do not include them here. The definitions of multiloop and external loop are generalized (Fig. 1(d)):

Multiloop: contains an external base pair $i.j$ and k tuples (i_1, j_1), (i_2, j_2), ..., (i_k, j_k), for some $k \geq 1$, along with the bases in $[i+1; j-1] - \cup [i_l; j_l], 1 \leq l \leq k$ all of which must be unpaired, where $i_l; j_l, 1 \leq l \leq k$, $i < i_1 < j_1 < i_2 < j_2 < ... < i_k < j_k < j$. Also, if $i_l.j_l$, $1 \leq l \leq k$, then k should be at least 2.

External loop: contains $k > 0$ tuples (i_1, j_1), (i_2, j_2), ..., (i_k, j_k) along with the bases in $[1; n] - \cup_{1 \leq l \leq k}[i_l; j_l]$, all of which must be unpaired, where $i_l; j_l$, $1 \leq l \leq k$, and $i_1 < j_1 < i_2 < j_2 < ... < i_k < j_k$.

We next introduce further types of elementary structures which are the consequence of having pseudoknotted base pairs and pseudoknotted regions.

Pseudoknotted loop: Let $[i; j']$ be a pseudoknotted closed region. Let the bands of $[i; j']$ be: $[i_1; i'_1] \cup [j'_1; j_1], [i_2; i'_2] \cup [j'_2; j_2], ..., [i_m; i'_m] \cup [j'_m; j_m]$. Let $[p_1; q_1]$, $[p_2; q_2]$, ..., $[p_k; q_k]$ be children of $[i; j']$ which are nested in $[i; j] - (\cup_{l=1}^{m}[i_l; i'_l] \cup_{l=1}^{m} [j'_l; j_l])$. The *pseudoknotted loop* corresponding to $[i; j']$ is the set: $\{(i_l, j_l), (i'_l, j'_l) | 1 \leq l \leq m\} \cup \{(p_l, q_l) | 1 \leq l \leq k\}$, along with the bases in: $[i; j'] - \cup_{l=1}^{k}[p_l; q_l] - \cup_{l=1}^{m}[i_l; i'_l] - \cup_{l=1}^{m}[j'_l; j_l]$ all of which must be unpaired (Fig. 1(e)).

Interior-pseudoknotted loop: contains two base pairs $i.j$ and $i'.j'$ where $i < i' < j' < j$, along with the bases in $[i+1, i'-1] \cup [j'+1, j-1]$ all of which must be unpaired. Moreover, there is a band $[bi; bi'] \cup [bj'; bj]$ such that $bi \leq i < bi'$ and $bj' < j \leq bj$. We refer to $i.j$ and $i'.j'$ as the interior-pseudoknotted loop external and internal base pairs respectively (Fig. 1(f)).

Multi-pseudoknotted loops: contains an external base pair $i.j$ and k tuples (i_1, j_1), (i_2, j_2), ..., (i_k, j_k), for some $k > 1$, along with the bases in $[i + 1; j - 1] - \cup_{1 \leq l \leq k}[i_l; j_l]$, all of which must be unpaired, where (i) there is a band $[bi; bi'] \cup [bj'; bj]$ such that $bi \leq i < bi'$ and $bj' < j \leq bj$, (ii) $i_l; j_l$, for all $1 \leq l \leq k$ except for exactly one tuple (i_{l_0}, j_{l_0}) for which $i_{l_0}; j_{l_0}$ is not true (i.e $[i_{l_0}; j_{l_0}]$ is not a closed region) and $i_{l_0}.j_{l_0}$ spans the band ($bi \leq i_{l_0} \leq bi'$ and $bj' \leq j_{l_0} \leq bj$), and (iii) $i < i_1 < j_1 < i_2 < j_2 < ... < i_k < j_k < j$ (Fig. 1(g)).

3.3 Enumerating Loops

We can enumerate the loops of a secondary structure in linear time. Each loop is fully specified by its list of tuples; thus an enumeration algorithm should list the tuples of each loop, with the external tuple first and the others in order.

Each node (closed region) of the tree corresponds to a hairpin loop, internal loop, multiloop, external loop, or pseudoknotted loop. A simple traversal of the tree suffices to enumerate such loops: when visiting a node, its closed region and the closed regions of its children (in order) are the needed tuples.

However, interior- and multi-pseudoknotted loops are not closed as their external base pair is pseudoknotted and spans a band. To enumerate these types of loops, two steps are needed:

Band finding: For each pseudoknotted closed region, construct the list of its bands regions, ordered by the left border index.

Loop finding: Identify all multi-pseudoknotted and interior-pseudoknotted loops, which are "nested" in the bands of the structure.

Algorithm 1 finds the bands of a pseudoknotted closed region $[i; j]$ of structure R. Loop finding is somewhat similar (details omitted).

Let L be a linked list representation of a secondary structure R for a strand of length n. In this representation, list elements are the base indices, with bidi-

rectional links between adjacent elements and additionally bidirectional links between paired indices. In this algorithm, $bp(i)$ denotes j if $i.j$ or $j.i$, and 0 if i is unpaired. Algorithm 1 takes as input a sublist BL of L starting from index i to index j, in which unpaired base indices, and base indices corresponding to nested closed regions, are removed. Sublist BL can be generated using the closed region tree in time proportional to the number of closed regions that are nested in $[i; j]$. Thus, BL is a linked list representation of spanning band base pairs in $i; j$. Inspired by Lemma 1, Algorithm 1 scans list BL from left to right to identify bands and their region's borders.

algorithm Band-Finding
input: BL, a linked list representation of spanning band base pairs in $[i; j]$
output: ordered linked list of band regions in $[i; j]$
1 $b_i := i$;
2 **repeat**
3 $b_j := bp(b_i)$; // $b_i.b_j$ is the outer closing pair of a band, B
4 $b'_i := b_i$;
5 $b'_j := b_j$;
7 **while** Next$(b'_i, BL) = bp($Prev$(b'_j, BL))$ **do**
8 $b'_i :=$ Next(b'_i, BL);
9 $b'_j :=$ Prev(b'_j, BL);
// $b'_i.b'_j$ is the inner closing pair of the band B So $B = [b_i; b'_i] \cup [b'_j; b_j]$ is a band of $i; j$
10 Add-Band-Region(BL, b_i, b'_i);
11 Add-Band-Region(BL, b'_j, b_j);
12 $b_i :=$Next-leftBase(b'_i, BL);
13 **until** $b_i = j + 1$;
14 **return** BL

Algorithm 1: Find bands of a pseudoknotted closed region.

Next(b'_i, BL) returns the index right after b'_i in BL and Prev(b'_j, BL) returns the index right before b'_j in BL. Next-leftBase(b'_i, BL) returns l, the first index after b'_i in BL for which $bp(l) > l$. $l.bp(l)$ will be the outer closing pair of the next band.

Add-Band-Region(BL, b, b') (i) replaces index b in BL with a list element containing the band region borders (b and b') and (ii) removes from BL all other base indices that lie within the region $[b; b']$. At the end, BL is an ordered list of band regions.

By traversing the closed regions tree and applying the above algorithm to each pseudoknotted closed region, all lists of band regions can be constructed in time linear in the number of base pairs in R.

4 Energy Model

In the standard thermodynamic model for pseudoknot free secondary structures, the energy of a loop is a function of (i) loop type, (ii) an ordered list of its base pairs or tuples, (iii) the bases forming each base pair, and (iv) the bases in the loop (if any) that are adjacent to each base pair. The energy of a secondary structure is then calculated by summing the free energy of its component loops.

For pseudoknotted structures, the standard thermodynamic model is extended so that the energy of a loop depends additionally on (v) the *location status* of the loop, which shows its position relative to pseudoknotted loops in the structure. The location status can be one of the following (Fig. 1(h)).

span-Band: Interior-pseudoknotted and multi- pseudoknotted loops are called span-Band loops, since their external base pair spans a band.

Each of the remaining loop types corresponds to a closed region. Suppose that such a loop, $\mathcal{L}$, with corresponding closed region $[i_{\mathcal{L}}; j_{\mathcal{L}}]$, is a child of pseudoknotted closed region $[i; j]$. Then $\mathcal{L}$ can have one of the following two location statuses:

in-Band: If $[i_{\mathcal{L}}; j_{\mathcal{L}}]$ is contained in a band region of $[i; j]$, then $\mathcal{L}$ is an in-Band loop.

out-Band: Otherwise $\mathcal{L}$ is an out-Band loop.

standard: Loops that are not of the three types above are called standard loops. Such loops do not span bands and are not children of pseudoknotted loops.

4.1 Energy Calculation

It is straightforward to extend the loop enumeration algorithm so that the loop's type and location status is output in addition to its list of tuples. For example, the type of a loop corresponding to a closed region can be determined from the number and types of its children (e.g. if the closed region is not pseudoknotted and has no children, it must be a hairpin loop; if it has one child which is not a pseudoknotted closed region then it must be an internal loop). The location status of a loop can be determined using additionally the ordered list of band regions of its parent (if any). Then the free energy of the structure can be calculated by adding up the free energy of all loops.

4.2 Discussion

In the Rivas-Eddy model [12], the energy of a loop is exactly as in the standard model (for pseudoknot free structures) if the loop does not span a band. The standard model is generalized in the case of multiloops, which may now contain pseudoknotted regions, as follows: the energy is of the form $a + bu + ch + dm$, where a, b, c, and d are constants independent of the loop, u is the number of unpaired bases of the loop, h is the number of tuples (i, j) of the multiloop with $i \cdot j \in R$, and m is the number of tuples (i, j) of the multiloop with $i \cdot j \notin R$.

For multi-pseudoknotted loops, the constants a, b, c, d are replaced by distinct constants a', b', c', d'. In contrast, in the D&P model [9], the energy of a multiloop and multi-pseudoknotted loop are calculated using the same constants. In both models, the energy of a pseudoloop is the sum of terms, with one term depending on the total number of unpaired bases, one term per tuple of the pseudoloop, and one term that depends on the location status of the pseudoloop; however the dependence on the location status is different for both models. An interesting direction for future work would be to establish which method is most biologically plausible (neither paper provides justification for their choice of model).

The notion of what is a multiloop in the Rivas-Eddy is perhaps unnaturally restrictive. An (artificially small) example lies in the structure $\{1 \cdot 4, 2 \cdot 9, 3 \cdot 5, 6 \cdot 8, 7 \cdot 10\}$. Here, the base pairs $2 \cdot 9, 3 \cdot 5$, and $6 \cdot 8$ could be considered to form a "multiloop", but it is not recognized as such by the Rivas-Eddy algorithm, and thus also not by our classification. (We note that the Dirks-Pierce model, being less general, does not handle such loops.) We expect that the Rivas-Eddy algorithm could be reformulated to assign multiloop energies to such loops.

5 Akutsu's Structure Class

Akutsu's dynamic programming algorithms for RNA secondary structure prediction handles a restricted class of pseudoknotted RNA structures, called secondary structures with recursive pseudoknots [1]. We present a concise characterization of the class of structures Akutsu's algorithm can handle.

In this section, we will represent secondary structures as patterns, in which information about unpaired bases and base indices is lost but the pattern of nesting or overlaps among base pairs is preserved. To define patterns precisely, we use ϵ to denote the empty string and N_n to denote the natural numbers between 1 and n (inclusive).

Patterns: A string P (of even length) over some alphabet Σ is a pattern, if every symbol of Σ occurs either exactly twice, or not at all, in P. We say that secondary structure R for a strand of length n corresponds to pattern P if there exists a mapping $m : N_n \to \Sigma \cup \{\epsilon\}$ with the following properties: (i) if $i.j \in R$ then $m(i) \in \Sigma$ and $m(i) = m(j)$, (ii) if $i.j$ and $j.i \notin R$ for all $j \in N_n$, then $m(i) = \epsilon$, and (iii) $P = m(1)m(2)...m(n)$.

We refer to the index of the first and the second occurrence of any symbol σ in P by $L(P, \sigma)$ and $R(P, \sigma)$ respectively (L for Left and R for Right). When P is understood, we use $L(\sigma)$ and $R(\sigma)$. For example, pattern $P = abccdebaed$ corresponds to the closed region $[21; 33]$ in Fig. 1, and $L(a) = 1$ and $R(a) = 8$.

In what follows, let P be a pattern of size $2n$ over an alphabet Σ of size n.

5.1 Definitions

Definition 1. Our definition: *P is a* simplest pseudoknot *if and only if either:*

B1: *$P = a_1 a_1$ (for some a_1), or*

B2: *Either* $P = a_1 a_i P_1 a_i a_1 P_2$ *or* $P = a_1 P_1 a_i a_1 a_i P_2$*, where* $a_1 P_1 a_1 P_2$ *is a* simplest pseudoknot.

P *is a* B&C simple pseudoknot *if and only if either it is a* simplest pseudoknot *or for some* a_1, a_i, $\ldots$ $a_r \in \Sigma$ *it is equal to* $a_1 P_1 a_1 a_i \ a_{i+1} \ldots a_r a_r \ldots a_{i+1} a_i P_2$*, where* $a_1 P_1 a_1 P_2$ *is a* simplest pseudoknot.

Theorem 1. *B&C simple pseudoknot is equivalent to Akutsu's definition of simple pseudoknot.*

Therefore, in what follows, we will simply refer to simple pseudoknots.

The following definition is derived from Akutsu [1].

Definition 2. *Pattern* P *is a* recursive pseudoknot *if and only if* P *is a* simple pseudoknot *or* $P = P_1 P_2 P_1'$ *where* P_2 *is a nonempty* simple pseudoknot *and* $P_1 P_1'$ *is a* recursive pseudoknot.

We say that an RNA secondary structure R *is a* secondary structure with recursive pseudoknots *or conveniently* recursive pseudoknot structure *if its corresponding pattern* P *is a* recursive pseudoknot.

Assume that C is the closed region corresponding to node V and $C_1, \ldots, C_m$ are the closed regions correspond to the children of V. Then we say that the pattern corresponding to C also corresponds to node V. Also, $C' = C - \cup_{i=1}^{m} C_i$ is called the *private region* corresponding to V and we refer to the pattern corresponding to C' as the *private pattern* of V.

Theorem 2. *R is an Akutsu (i.e. recursive pseudoknot) structure if and only if all of the private patterns corresponding to the nodes in $T(R)$ are simple pseudoknots.*

5.2 Akutsu Tests

Our algorithm for testing whether a pattern P is a simple pseudoknot has two steps. In the first step it deals with the $a_i a_{i+1} \ldots a_r a_r \ldots a_{i+1} a_i$ subpattern and removes it from P, making the pattern a *simplest pseudoknot*. This can be done in linear time by scanning the symbols of P, starting from the symbol after the second occurrence of a_1, and removing the subpattern $a_i a_{i+1} \ldots a_r a_r \ldots a_{i+1} a_i$ if any.

Next the algorithm determines if P is a simplest pseudoknot, building on both cases in the definition of simplest pseudoknot. We define two *simplify* operations according to **B2**: (**i**) $a_1 a_i S_1 a_i a_1 S_2$ is converted to $a_1 S_1 a_1 S_2$, and (**ii**) $a_1 S_1 a_i a_1 a_i S_2$ is converted to $a_1 S_1 a_1 S_2$. We define one more operation, *final* operation, according to **B1**: (**iii**) $a_1 a_1$ is converted to ϵ. In these cases we say that a simple/final operation is applicable to a_1.

The linear time algorithm for testing whether the pattern P is a simplest pseudoknot (1) applies one of the simplify operations, **i** or **ii**, on the first symbol, a_1, if applicable, repeatedly (2) does the *final* operation, **iii**, on a_1 if it is applicable. (3) return true if the pattern is empty and false otherwise.

Thus, using Theorem 2, to test whether a secondary structure R is an Akutsu (i.e. recursive pseudoknot) structure, it is sufficient to check whether the private pattern corresponding to each node of $T(R)$ is a simple pseudoknot. It is straightforward to generate the private pattern for all nodes in linear time; thus the overall algorithm is a linear time algorithm.

5.3 Classification of Biological Structures

Condon et al. [7] provide linear time algorithms to test if an input structure is in the R&E and D&P classes. To compare the generality of Akutus's algorithm with those of R&E and D&P, we applied our algorithms for membership in Akutsu's recursive class along with those of Condon et al.[7] to classify biological structures from several sources [3, 10, 6, 4, 5, 13, 15]. As results show (Table 1), exactly 2 of the structures are in class A but not in class D&P.

6 Conclusions

In this work we present a precise definition of the structural elements in a secondary structure, and a comprehensive way to classify the type of loops that arise in pseudoknotted structure. Based on an algorithm of Bader et al. [2], we also introduced a linear time algorithm to parse a pseudoknotted secondary structure to its component loops, and to calculate its the free energy. Finally, we applied our algorithm to compare the generality of Akutsu's algorithm with those of Dirks and Pierce and Rivas and Eddy on a large test set of biological structures.

Our work can be continued in future in several directions. First, heuristic algorithms commonly use a procedure to calculate the free energy for a given sequence and structure. Incorporating our linear time free energy calculation algorithm into heuristic algorithms may cause improvements in their efficiency. Second, it would be interesting to investigate the structures which are in Akutsu's class but not in D&P class. Third, there is no linear time characterization of Uemura's [14] algorithm and having one makes it possible to figure out about the differences between Uemura's class of structures and other classes of structures (A, D&P, and R&E). Fourth, the parsing algorithm can be used to analyse known biological RNA structures, in order to find out what structures occur more frequently in biology. Finally, it would be useful to refine the thermodynamic model presented in this paper, to obtain mfe predictions of better quality.

Acknowledgement: We would like to thank Satoshi Kobyashi for his useful comments and pointing out an error in an earlier version of the paper.

References

1. Akutsu, T.: Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics* **104** (2000) 45–62.

	PBase	Pseudo Viewer	Gutell	RCSB	RNase	SR PDB	tm RNA
# Strs	240	15	426	279	468	4	7
Avg. #Bps	14.2	144	970.6	35.5	198.4	92	85.71
D&P	232	11	354	244	95	3	5
A	232	11	354	246	95	3	5
R&E	240	15	369	274	468	4	7

Table 1. Structure classification. Columns 2-8 present data for each RNA data set. For each data set (column), the entry in the first row lists the number of structures in the data set. The second row lists the average number of base pairs in the structures. The remaining rows list the number of structures of the data set that are in D&P, A, and R&E classes.

2. Bader, D. A., Moret, B. M.E., Yan, M.: A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *Journal of Computational Biology* **8** (2001) 483–491.
3. Batenburg, F. H. D. van et al.: Pseudobase: a database with RNA pseudoknots. *Nucl. Acids Res.* **28** (2000) 201–204.
4. Berman, H.M. et al.: The Nucleic Acid Database; A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys. J.* **63** (1992) 751–759.
5. Brown, J.W.: The Ribonuclease P Database. *Nucl. Acids Res.* **27** (1999) 314.
6. Cannone, J.J. et al.: The Comparative RNA Web (CRW) Site; an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3** (2002)
7. Condon, A., Davy, B., Rastegari, B., Zhao, S. and Tarrant, T.: Classifying RNA pseudoknotted structures. *Theor. Comput. Sci.* **320** (2004) 35–50.
8. Dennis, C.: The brave new world of RNA. *Nature* **418** (2002) 122–124.
9. Dirks, R. M., Pierce, N. A.: A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* **24** (2003) 1664–1677.
10. Han, K., Byun, Y.: PseudoViewer2: visualization of RNA pseudoknots of any type, *Nucl. Acids Res.* **31** (2003) 3432–3440.
11. Lyngsø R. B., Pedersen, C. N.: RNA pseudoknot prediction in energy-based models. *J. Computational Biology* **7** (2000) 409–427.
12. Rivas, E., Eddy, S. R.: A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Molecular Biology* **285** (1999) 2053–2068.
13. Rosenblad, M.A., Gorodkin, J., Knudsen, B., Zwieb, C., Samuelsson, T.: SRPDB: Signal Recognition Particle Database. *Nucl. Acids Res.* **31** (2003) 363–364.
14. Uemura, Y., Hasegawa, A., Kobayashi, S., Yokomori, T.: Tree adjoining grammars for RNA structure prediction. *Theor. Comput. Sci.* **210** (1999) 277–303.
15. Zwieb, C., Gorodkin, J., Knudsen, B., Burks, J., Wower, J.: tmRDB (tmRNA database). *Nucl. Acids Res.* **31** (2003) 446-447.