Rapid *ab initio* RNA Folding Including Pseudoknots via Graph Tree Decomposition

Jizhen Zhao[1], Russell L. Malmberg[2], and Liming Cai[1]

[1] Department of Computer Science, University of Georgia, Athens, GA 30602, USA.
jizhen, cai@cs.uga.edu *
[2] Department of Plant Biology, University of Georgia, Athens, GA 30602, USA.
russell@plantbio.uga.edu

Abstract. The prediction of RNA secondary structure including pseudoknots remains a challenge due to the intractable computation of the sequence conformation from intriguing nucleotide interactions. Optimal algorithms often assume a restricted class for the predicted RNA structures and yet still require a high-degree polynomial time complexity, which is too expensive to use. Heuristic methods may yield time-efficient algorithms but they do not guarantee optimality of the predicted structure. This paper introduces a new and efficient algorithm for the prediction of RNA structure with pseudoknots for which the structure is not restricted. Novel prediction techniques are developed based on graph tree decomposition. In particular, stem overlapping relationships are defined with a graph, in which a specialized maximum independent set (IS) corresponds to the desired optimal structure. Such a graph is tree decomposable; dynamic programming over a tree decomposition of the graph leads to an efficient algorithm. The new algorithm is evaluated on a large number of RNA sequence sets taken from diverse resources. It demonstrates overall sensitivity and specificity that outperforms or is comparable with those of previous optimal and heuristic algorithms yet it requires significantly less time than other optimal algorithms.

Keywords: RNA secondary structure prediction, pseudoknot, thermodynamic energy, tree decomposition, tree width, maximum independent set

1 Introduction

The secondary structure of an RNA molecule is formed due to short or long distance pairings between nucleotides in the sequence. Base pair regions either single, nested or parallel are called *stem-loops*; base pair regions crossing each other are called *pseudoknots* [23]. Pseudoknots are important structures in RNA molecules and often play important functional roles [12] such as catalysis, RNA splicing, transcription regulation. Knowing the secondary structures of RNA molecules is critical for determining their three dimensional structures and understanding their functions. Automated prediction of RNA secondary structure

* To whom correspondence should be addressed.

is thus in demand since it is expensive and time consuming to experimentally determine the structure.

It is computationally challenging to predict RNA secondary structure including pseudoknots. In particular, the problem of predicting RNA pseudoknots with the minimum free energy is provably NP-hard [13]. Practical approaches to cope with this computational challenge are either to restrict the class of pseudoknots under consideration or to employ heuristics in the algorithms. Optimal algorithms for restricted pseudoknot classes are usually thermodynamics-based, extended from Zuker's algorithm for the prediction of pseudoknot-free structures [25]. In such algorithms, the predicted optimal structure of a single RNA sequence is the one with the global minimum free energy based on a set of experimentally determined parameters. Among these algorithms, PKNOTS [17] can handle the widest classes of pseudoknots. However, its time complexity $O(n^6)$ makes it infeasible to fold RNA sequences of a moderate length. The computation efficiency may be improved at the cost of further restricting the structure of pseudoknots [16], but still with a time complexity $O(n^5)$ or $O(n^4)$. Most such algorithms produce only the optimal solution, while suboptimal ones that may reveal the true structure are often ignored.

On the other hand, computationally efficient heuristic methods have also been explored to allow unrestricted pseudoknot structures. Iterated loop matching (ILM) [19] is one such method. It finds the most stable stem, adds it to the candidate secondary structure and then masks off the bases forming the stem and iterates on the left sequence segments until no other stable stem can be found. One structure is reported at the end. Another algorithm, HotKnots [16], does the prediction in a slightly different way. It keeps multiple candidate structures rather than only one and builds each of them in a similar but more elaborate way. These methods can usually be fast, yet they often do not provide an optimality guarantee for the predicted structure or a quality measure on the predicted structure with respect to the optimal structure. Other heuristic methods based on genetic algorithms and Monte Carlo simulation usually do not address the optimality issue either [1, 5].

In this paper, we introduce a novel approach for the optimal prediction of RNA pseudoknots for which the structure is not restricted. Our method is based on a simplified thermodynamic model without accounting for loop energies [15, 19]. In this method, stable stems are selected from an RNA sequence as vertices of a graph; vertices are connected with edges if corresponding stems conflict (i.e., overlap) in their positions in the sequence. The optimal structure of an RNA sequence corresponds to a collection of non-conflicting stable stems, which can be found by seeking the maximum weighted independent set (WIS) from the graph. We observe that stable stems can be so selected that the resulting graph is of a moderately small tree width t. Based on a tree decomposition of the graph, a dynamic programming algorithm for WIS of the worst-case time complexity $O(1.44^t n)$ is obtained, where n is the number of vertices in the graph, at most quadratic in the length of the RNA sequence. This is an efficient prediction algorithm parameterized on the tree width t, which is usually small.

We implemented our algorithm TdFOLD and evaluated its performance on various RNA sequence sets from different sources. The test results showed high efficiency and high accuracy for our algorithm. TdFOLD was tested against PKNOTS, ILM and HotKnots on a set of 50 tRNA's, a set of 50 small RNA sequences containing pseudoknots with length ranging from 23 to 113, and a set of 11 large RNA's with length range from 210 to 412. The results showed that overall, in terms of the sensitivity and specificity of the prediction, TdFOLD outperforms the optimal algorithm PKNOTS and the heuristic algorithms ILM and HotKnots. In time efficiency, it outperforms PKNOTS and HotKnots, and is comparable with ILM. Our algorithm will also output suboptimal structures without spending much more time than reporting the optimal structure.

Graph theoretic methods have previously been explored for RNA structure prediction [23]. Our method is different from the previous ones in two respects. Our graphs constructed from the RNA sequence contain vertices describing stems instead of nucleotides; making the stem to be the smallest structural unit can greatly simplify the complexity of the problem. More importantly, our graph algorithm takes advantage of the tree decomposition technique on the formulated graphs. In fact, it has been demonstrated that the RNA secondary structure can be profiled with a conformational graph of small tree width [21]. The underlying graph constructed for the *ab initio* structure prediction is essentially an augmentation of the conformational graph in which additional vertices and edges are added only for the overlapping stems, thus inheriting the tree decomposability which makes the algorithm efficient.

2 Methods and Algorithm

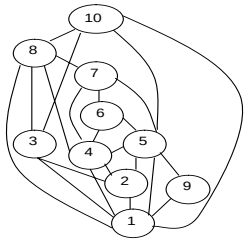
Given an RNA sequence, our algorithm first builds a pool of stable stems, then finds a number of secondary structures with (near) minimum total stem energies by a tree decomposition based procedure for a graph formed by the stable stems. These predicted secondary structures are then reordered by counting the stem and loop energies together.

2.1 Problem formulation

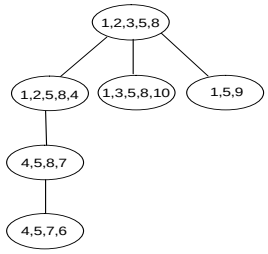
A (canonical) base pair is either a Watson-Crick pair (A-U or C-G) or for wobble pair G-U. A *stem* is a set of stacked nucleotide base pairs on an RNA sequence s. In general a stem S can be associated with four positions (i^l, j^l, i^r, j^r), where $i^l < j^l < i^r < j^r$, on the sequence s such that (a) $(s[i^l], s[j^r])$ and $(s[j^l], s[i^r])$ are two canonical base pairs; and (b) for any two base pairs $(s[x], s[y]), (s[z], s[w])$ in the stem S, either $i^l \leq x < z \leq j^l$ and $i^r \leq w < y \leq j^r$, or $i^l \leq z < x \leq j^l$ and $i^r \leq y < w \leq j^r$. Region $s[i^l..j^l]$ is the *left region* of the stem and $s[i^r..j^r]$ is the *right region* of the stem. Stem S is *stable* if the formation of its base pairs allows the thermodynamic energy $\Delta(S)$ of the stem to be below a predefined threshold parameter $E < 0$. Figure 1(a) shows all the stable stems in Ec_Pk_4

ID	L	R	En
1	1..10	27..36	-12.4
2	5..10	47..52	-7.8
3	1..7	34..40	-6.8
4	9..12	22..25	-6.2
5	10..13	27..30	-5.5
6	11..14	19..22	-5.5
7	12..14	23..25	-5.4
8	23..25	35..37	-5.4
9	15..18	30..33	-5.3
10	27..29	34..36	-5.1

(a)



(b)



(c)

Fig. 1. (a) Ten stable stems in Ec_Pk_4, the fourth pseudoknot in E.coli tmRNA molecule, including their left and right regions, and thermodynamic energies; (b) stem graph for Ec_Pk_4; and (c) a tree decomposition of the stem graph with tree width 4.

with $E = -5$ kcal/mol, the fourth pseudoknot in E.coli tmRNA [24], and their corresponding free energy values.

A *stem graph* $G = (V, E)$ can be defined for the RNA sequence s, where each vertex in V uniquely represents a stable stem on s, and E contains an edge between two vertices if and only if the corresponding two stems (a, b, c, d) and (x, y, z, w) conflict in their positions, i.e., one or both of the regions $s[a..b]$ and $s[c..d]$ overlap with at least one of the regions $s[x..y]$ and $s[z..w]$. Figure 1(b) shows the stem graph for Ec_Pk_4 constructed according to the stable stems given in Figure 1(a). The stem graph is a weighted graph, with a weight on every vertex. Usually, the weight of a vertex can simply be the absolute value of the thermodynamic energy $\Delta(S)$ of the stem S corresponding to the vertex. The weight may also be adjusted by scaling it (non-)linearly according to the length of the corresponding stem or the distance between the left and right regions of the stem. The problem of predicting the optimal structure of the RNA then corresponds to finding a collection of non-conflicting stems from its stem graph which achieves the maximum total weight. This is exactly the same as the graph theoretic problem: finding the maximum WIS in the stem graph. Note the weight for an IS representing a secondary structure is based on the total energies of the stems only (similar models were previously adopted by both primitive method [15] and more elaborate one [19]).

2.2 Identifying stable stems

For our purpose, stable stems are defined according to a set of parameters. In particular, a stem contains at least P base pairs; the loop length in between the left and right region of the stem is at least L; the thermodynamic energy is at most E. Bulges within a stem are allowed, for which the stem essentially becomes a set of substems separated by the bulges. In addition, parameter T limits the minimum substem length, and parameter B limits the maximum bulge length. The thermodynamic energy $\Delta(S)$ of stem S is calculated by taking into account

both the stacking energies and the destabilizing energies caused by bulges. A procedure similar to the one used in [11] is employed to identify all the stable stems. These stable stem pool can be extended by introducing maximal substems that can resolve the confliction and meet the requirements defined by the above parameters for each pair of overlaped stems in the pool.

2.3 Tree decomposition based algorithm

Definition [18] A *tree decomposition* of graph $G = (V, E)$ is a pair (T, X) if it satisfies:

1. $T = (I, F)$ is a tree with node set I and edge set F,
2. $X = \{X_i : i \in I, X_i \subseteq V\}$, $\bigcup_i X_i = V$ and $\forall u \in V$, $\exists i \in I$ such that $u \in X_i$,
3. $\forall (u, v) \in E$, $\exists i \in I$ such that $u, v \in X_i$,
4. $\forall i, j, k \in I$, if k is on the path that connects i and j in tree T, $X_i \cap X_j \subseteq X_k$

The *width* of a tree decomposition (T, X) is $\max_{i \in I} |X_i| - 1$. The *tree width* of the graph G is the minimum tree width over all possible tree decomposition of G. If T is restricted to be a path, we refer (T, X) as a *path decomposition* and the best width over all of the path decompositions as the *path width* of G. The tree decomposition is rooted in the deep graph minor theorems by Robertson and Seymour [18]. It provides a topological view on a graph and the tree width measures how much the graph is "tree-like". Figure 1(c) shows a tree decomposition for the stem graph given in Figure 1(b).

Many computationally intractable graph problems can be easily solved on graphs of small tree width. In particular, a large number of such graph problems, while intractable on general graphs, can be solved in linear time, given a tree decomposition of tree width $\leq t$, for a fixed t. Maximum WIS is one such problem [3]; it has time complexity $O(2^t n)$. For the RNA stem graphs, we observe that vertices contained in every node of a tree decomposition can be partitioned into a small collection of maximal cliques, thus the factor 2^t can be further reduced. For example, in Figure 1(c), node $\{1, 2, 5, 3, 8\}$ contains two cliques $\{1, 2, 5\}$, and $\{3, 8\}$ (also see Figure 1(b)). In general, let $C_1, \ldots, C_q$, where $\sum_{i=1}^{q} |C_i| = t$, be the maximal cliques contained in a node, for some small q, then the number of valid partial ISs for the vertices in the node is at most $\prod_{i=1}^{q} |C_i| \leq (t/q)^q$. While the right term may reach the worst case extreme $e^{t/e} \approx 2^{0.53t}$ when $t/q = e$, the base of natural logarithm, in reality, the worst case may never occur because q usually is small. In the above example, the factor is reduced to $3 \times 2 = 6$ in contrast to the number $2^5 = 32$.

Algorithm details Now we describe the tree decomposition based dynamic programming algorithm that finds the maximum WIS from the stem graph $G = (V, E)$. It assumes a binary tree decomposition (T, X), where $X = \cup X_{i=1}^{m}$, for the stem graph, where $m = O(|V|)$, $|X_i| = t$, for $i = 1, \ldots, m$. We only discuss the process for achieving the optimal solution. The technical details for getting suboptimal solutions are similar.

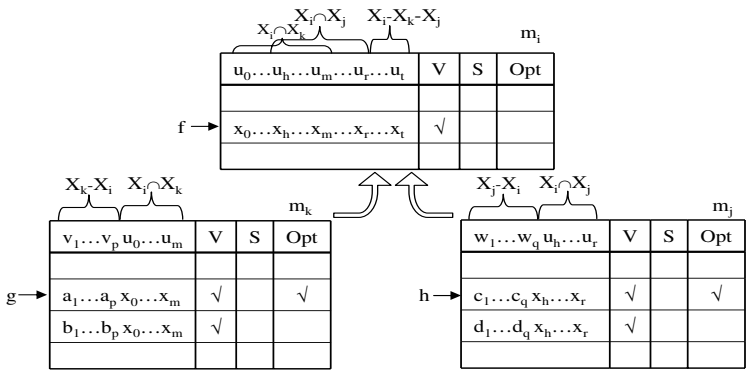


Fig. 2. Dynamic programming table construction over tree decomposition. Table m_i is computed also based on the computed tables m_k and m_j. Row $f = (x_0, \ldots, x_h, \ldots, x_m, \ldots, x_r, \ldots, x_t)$ in table m_i is computed from row g in table m_k and row h of table m_j, Row g is the optimal for columns $X_k - X_i$ given the value $(x_0, \ldots, x_m)$ for columns $X_k \cap X_i$. Similarly, row h is the optimal for columns $X_j - X_i$ given the value $(x_h, \ldots, x_r)$ for columns $X_j \cap X_i$.

The algorithm constructs one dynamic programming table m_i for every tree node $X_i = \{v_1, \ldots, v_t\}$. Table m_i records all possible partial ISs in the subgraph induced by the set of all the vertices in the subtree rooted at i of the tree decomposition. There are t columns in the table m_i, one for each vertex in the corresponding tree node X_i. Rows are the combinations of these vertices; a vertex is selected if and only if the corresponding column takes value 1. There are additional three columns V, S, Opt in the table.

These tables are constructed in a bottom-up fashion, from leaves to the apex of the tree decomposition (see Figure 2). Each row of a table is a combination of the vertices in the corresponding node. Column V is set 1 if the row represents a valid IS. For a leaf node, S is 0 if the row is not a valid IS; otherwise S is the corresponding weight of the set. For an internal node i that has two children j and k whose tables m_j, m_k have been computed, for each row in table m_i, column S is computed as $S = w_1 + w_2 + w_3 - w_4$, where

- w_1 is the weight of the row in table m_j with the same combination in the columns corresponding to the vertices in $X_j \cap X_i$ that has column $Opt = 1$;
- w_2 is the weight of the row in table m_k with the same combination in the columns corresponding to the vertices in $X_k \cap X_i$ that has column $Opt = 1$;
- w_3 is the weight of the IS formed by the choices in columns corresponding to the vertices in $X_i - X_j - X_k$; and
- w_4 is the weight of the IS formed by the same combination in the columns corresponding to the vertices in $X_i \cap X_j \cap X_k$.

Column Opt is set 1 if and only if the row represents a valid IS and S in this row is optimal among all the rows with different choices in the columns corresponding

to the vertices in $X_i - X_p$ given the chosen values same as this row in the columns corresponding to the vertices in $X_i \cap X_p$, where node p is the parent of node i.

As mentioned earlier, the enumeration of the combinations of the graph vertices in tree node X_i is along a number of maximal cliques. In general, a greedy algorithm is used to partition set X_i into a collection of cliques. Consider the sequence as a straight line and the left (right) region of a stem as an interval. Let all the left regions of the stable stems included in the tree node form an interval graph. Choose an interval (left region) with the right end at the left most position among all of the intervals, record all the intervals overlap with this interval as a clique and remove them, recursively call on the interval graph left until it is empty. A linear time in t is enough for this procedure.

Tree decomposition of stem graph Finding the optimal tree decomposition is NP-hard [2], we use a simple, fast heuristic algorithm to produce a tree decomposition for the given stem graph. This algorithm is based on a heuristic method for greedy fill-in [10]. This method will produce a tree decomposition with small tree width but not necessary the optimal one.

Reordering suboptimal structures The list of candidate structures, including the optimal and the suboptimal ones, are reordered based on a more sophisticated energy model. In particular, we recalculate the free energy for each of the candidate structures using a procedure implemented in [16] according to the energy model in [20, 14] combined with the one in [6], which take the stem stabilizing energies, loop destabilizing, and pseudoknot energies into account.

3 Evaluation Results

3.1 Data sets and experiment details

We used three sets of RNA sequences to evalute the algorithm (see Table 1). The first set is 50 tRNAs with lengths ranging from 71 to 79 (with the average 75). The second set is 50 small RNA sequences or sequence segments with pseudoknot structures of lengths ranging from 23 to 113 (with the average 53). The third set is 11 large RNA sequences of lengths ranging from 210 to 412 (with the average 344).

We compared the performance of our algorithm TdFOLD and that of algorithms PKNOTS [17], ILM [19], and HotKnots [16]. We ran all these algorithms on the tRNAs and the set of small pseudoknot RNAs, and run all but PKNOTS on the set of large RNAs. We evaluated both accuracy and efficiency of these algorithms. The accuracy is measured in both sensitivity and specificity. Let RP be the number of base pairs in the real structure, TP (true positive) be the number of correctly predicted base pairs and FP (false positive) be the number of predicted base pairs that do not exist as real structures. We define SE

(sensitivity) as TP/RP, and SP (specificity) as $TP/(TP + FP)$. The perfect prediction should yield 1 for both sensitivity and specificity values.

For tRNA, we turned off the pseudoknot option for PKNOTS since we already know they are pseudoknot free. For TdFOLD, parameters were set to default values and the number of output solutions was set to 40 for tRNAs and small pseudoknotted RNAs. The parameters were adjusted for each of the large sequences. The experiments were run on a PC with 2.8 GHz Intel(R) Pentium 4 processor and 1-GB RAM, running RedHat Enterprise Linux version 4 AS.

Set one: tRNA[22]	
GA0001 GA1262 GA2492 GA3755 GA4966 GC2866 GD1723 GD5199 GE2095 GE4739 GF1407 GF4687 GG0841 GG2136 GG3917 GH0128 GH4536 GI1748 GI4502 GK1078 GK4537 GM0313 GM2284 GM4471 GM5945 GN2837 GP1341 GP3879 GP5312 GQ2684 GR0044 GR0793 GR1516 GR2309 GR3541 GR4508 GR4705 GR4740 GR5278 GT0109 GT1418 GT4178 GT5273 GV0579 GV1734 GV4391 GV5554 GW1796 GW5332 GY4135	
Set two: small RNAs	
Sequence type	Sequence IDs
aptamers	NGF-L6 [24]
antizyme ribosomal frame shifting site	Rr_ODCanti [24]
HIV-1-RT ligand RNA	HIVRT32, HIVRT322, HIVRT33 [16]
hepatitis virus ribozyme	HDV, HDV_anti [16]
mRNA	Bt-PrP, Ec_alpha, Ec_S15, Hs-PrP, T4_gene32 [24]
rRNA	Sc_18S-PKE21-7 [24]
ribozymes	HDV-It_ag [24]
ribozymes	satRPV, Tt-LSU-P3P7, Bp_PK2 [24]
tmRNA	Lp_PK1, Ec_PK1, Ec_PK4 [24]
telomerase RNA	T.the_telo [24]
viral tRNA like	OYMV, APLV, CGMMV, SBWMV1, BSMVbeta, CGMMV_PKbulge, ORSV-S1, AMV3 [24]
viral 3'UTR	TMV-L_UPD-PK3, STMV_UPD1-PK3, BVQ3_UPD-PKb, BSBV1_ , PSLVbeta_UPD-PK1, PSLVbeta_UPD-PK3, BSBV3, UPD-PKc, SBWMV1_UPD-PKb [24]
viral ribosomal RNA shifting signals	EIAV, PLRV-S [24]; minimal IBV, MMTV, MMTV-vpk, pKA-A, BWYV, SRV-1, T2_gene32[9]
viral RNA	PSIV_IRES [24]; TYMV, TMV.L, TMV.R [16]
Set three: large RNAs	
Sequence type	Sequence IDs
RNaseP RNA	A.ferr, A.laid (pseudoknot free), A.tum, B.anth, B.halo,
	CPB147, D.desu, EM14b-9, E.ther, T.rose [4]
telomerase RNA	telo.human [5]

Table 1. Test sets: sequence IDs with their reference citations.

3.2 Testing results

Tables 2 summarize the testing results for different programs on the three RNA data sets. It shows that TdFOLD has sensitivity 0.81 and specificity 0.75 on average for the tRNA prediction, which are slightly better than PKNOTS and significantly better than ILM and HotKnots. For the small pseudoknotted RNAs, TdFOLD has average sensitivity 0.76, which is less than PKNOTS but greater than ILM and HotKnots. On the other hand, TdFOLD has average specificity 0.79, which outperforms all the others. TdFOLD is slightly better in overall accuracy than PKNOTS, which reports the optimal structure according to its sophisticated energy model. This suggests that considering the stems as prediction units can filter some noise. For the large RNA's, TdFOLD maintains the same sensitivity (0.54) as HotKnots, which is slightly better than ILM. TdFOLD has the highest specificity on average.

		TdFOLD			HotKnots			ILM			PKNOTS		
		SE	SP	T	SE	SP	T	SE	SP	T	SE	SP	T
tRNA	min	0.33	0.29	0.26	0.33	0.25	0.57	0.33	0.25	0.01	0	0	0.11
	max	1.00	1.00	1.37	1.00	1.00	8.32	1.00	1.00	0.15	1.00	1.00	0.24
	average	0.81	0.75	0.54	0.72	0.66	3.33	0.75	0.61	0.03	0.78	0.73	0.41
small	min	0	0	0.04	0	0	0.05	0	0.25	0.001	0	0	0.27
	max	1.00	1.00	0.57	1.00	1.00	57.0	1.00	1.00	0.05	1.00	1.00	>1hr
	average	0.76	0.79	0.36	0.69	0.72	5.84	0.73	0.69	0.03	0.78	0.73	1066
large	min	0.18	0.17	0.46	0.24	0.18	157	0.38	0.25	0.71			
	max	0.86	0.73	14.5	0.68	0.63	29710	0.77	0.82	1.49			
	average	0.54	0.53	3.97	0.54	0.49	4456	0.51	0.44	0.97			

Table 2. Summary of testing results on tRNAs, smale and large RNAs, where SE: sensitivity, SP: specificity, T: time (in seconds, if not otherwise noted).

Efficiency comparisons are also given in Tables 2 on each data set, respectively. For tRNA's, the average running time of 0.54 seconds for TdFOLD is slower than the average 0.03 of ILM and the average 0.41 of PKNOTS but faster than the average 3.33 of HotKnots. This is not a surprise because we turned the pseudoknot option off for PKNOTS. For small pseudoknotted RNA's, TdFOLD is slower than ILM (0.36 vs. 0.03 seconds), while much faster than HotKnots and PKNOTS (5.84 and 1066 seconds). For large RNA sequences, it is comparable (slightly slower) than ILM (3.97 vs. 0.97 seconds) while much faster than HotKnots (4456 seconds) on average. In general, the speed of TdFOLD is comparable to ILM and much faster than PKNOTS and HotKnots.

According to Table 2, all of the programs could predict some sequences (different for each program) totally wrong (zero sensitivity and/or specificity). This reveals that the available thermodynamic parameters for RNA secondary structures may not be optimal for all RNA classes. Thus it is hard to guarantee that the structure with the minimum free energy is the true structure. This makes

the output of a list of low energy suboptimal structures a valuable feature of a structure prediction algorithm. The prediction results of TdFOLD for 23 tRNAs and 19 short pseudoknotted RNAs are improved by considering the top five structures, rather than only the top one among the 40 output predictions for each sequence. By "improved" we mean that there is at least one suboptimal prediction with both the sensitivity and specificity better than (or the same as) those of the optimal prediction. If there is more than one prediction improved over the top one, we choose the best among all the improved. For example, the average sensitivity and specificity are improved to 0.91 and 0.85 for tested tRNAs, 0.81 and 0.85 for tested short pseudoknotted RNAs.

4 Discussion and conclusion

When related structurally homologous sequences are available, the accuracy of RNA structure prediction can usually be improved through the use of comparative analysis. A fully automated comparative analysis process exists [8, 7] for consensus structure prediction of pseudoknot free RNAs, which iterates between the following two steps: (a) build an optimal (or nearly optimal) structure model given the current multiple alignment; and (b) build a multiple alignment given the current structure model. Nevertheless, for RNA pseudoknots, both algorithms for step (a) and (b) can be computationally intensive; the implementation remains a computational challenge.

The tree decomposable model and tree decomposition based techniques make it possible to implement efficiently the automated comparative analysis process. Based on an earlier work of ours, pseudoknots can be profiled with the conformational graph model [21] of small tree width; the efficient optimal structure-sequence alignment developed is ideal for step (b). In addition, the algorithm introduced in this paper can be employed for step (a), to construct a structure model for multiple RNAs. As it was done for pseudoknot-free RNAs, the mutual information content $M_{i,j}$ can be computed for every pair of aligned columns i, j, which is defined as the relative entropy

$$M_{i,j} = \sum_{x_i, y_j \in \{A,C,G,U\}} f(x_i, y_j) \log \frac{f(x_i, y_i)}{f(x_i) f(y_j)}$$

where $f(x_i, y_j)$ is the frequency for nucleotides x_i, y_j to occur in pair in these two columns i, j, and $f(x_i)$ and $f(y_j)$ are for independent occurrences. The multiple alignment can be regarded as a "generic sequence" consisting of columns as "nucleotides". The pairwise interactions between columns result in a conformation structure of the "generic sequence", yielding a consensus structure for the multiple sequences. Therefore, we can use our structure prediction algorithm TdFOLD to predict the structure of the "generic sequence" using the mutual information content $M_{i,j}$ as "pairing energy" between columns i and j.

In conclusion, in this paper, we presented a tree decomposition based fast RNA folding algorithm, which is efficient, accurate, not limited to any specific

class of pseudoknots, and can report a list of suboptimal structures. Combined with an efficient structure-sequence alignment algorithm we developed earlier [21], it also can be used to implement an automated comparative RNA structure analysis process that can infer the pseudoknot consensus structure from a set of unaligned RNA sequences.

Acknowledgment

This work was supported in part by the NIH BISTI grant No: R01GM072080-01A1.

References

1. J. Abrahams, M. van den Berg, E. van Batenburg, and C. Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res.*, 18:3035–3044, 1990.
2. H. L. Bodlaender. Classes of graphs with bounded tree-width. *Tech. Rep. RUU-CS-86-22, Dept. of Computer Science, Utrecht University, the Netherlands*, 1986.
3. H. L. Bodlaender. Dynamic programming algorithms on graphs with bounded tree-width. In *Proceedings of the 15th International Colloquium on Automata, Languages and Programming*, pages 105–119. Springer Verlag, Lecture Notes in Computer Science, vol. 317, 1987.
4. J. Brown. The ribonuclease p database. *Nucleic Acids Res.*, 27:314, 1999.
5. J.-H. Chen, S.-Y. Le, and J. V. Maize. Prediction of common secondary structures of RNAs: a genetic algorithm approach. *Nucleic Acids Research*, 28(4):991–999, 2000.
6. R. Dirks and N. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, 24:16641677, 2003.
7. R. Durbin, S. R. Eddy, A. Krogh, and G. J. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
8. S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22:2079–2088, 1994.
9. D. Giedroc, C. Theimer, and P. Nixon. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frame shifting. *Journal of Molecular Biology*, 298:167–185, 2000.
10. I. V. Hicks, A. M. C. A. Koster, and E. Kolotoglu. Branch and tree decomposition techniques for discrete optimization. In *Tutorials in Operations Research: INFORMS – New Orleans 2005*. 2005.
11. Y. Ji, X. Xu, and G. D. Stormo. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, 20(10):1591–1602, 2004.
12. A. Ke, K. Zhou, F. Ding, J. H. Cate, and J. A. Doudna. A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature*, 429:201205, 2004.
13. R. B. Lyngso and C. N. S. Pedersen. RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology*, 7(3-4):409–427, 2000.

14. D. H. Mathews, J. Sabina, M. Zuker, and C. N. S. Pederson. Expanded sequence dependence of the thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.
15. R. Nussinov, G. Pieczenik, J. Griggs, and D. Kleitman. Algorithms for loop matchings. *SIAM Journal of Applied Mathematics*, 35:68–82, 1978.
16. J. Ren, B. Rastegart, A. Condon, and H. H. Hoos. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 11:1194–1504, 2005.
17. E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285:2053–2068, 1999.
18. N. Robertson and P. D. Seymour. Graph minors ii. algorithmic aspects of tree width. *Journal of Algorithms*, 7:309–322, 1986.
19. J. Ruan, G. D. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20(1):58–66, 2004.
20. M. J. Serra, D. H. Turner, and S. M. Freier. Predicting thermodynamic properties of RNA. *Meth. Enzymol.*, 259:243–261, 1995.
21. Y. Song, C. Liu, R. L. Malmberg, F. Pan, and L. Cai. Tree decomposition based fast search of RNA structures including pseudoknots in genomes. In *Proceedings of 2005 Computational System Bioinformatics Conference*, pages 223–234. IEEE Computer Society, 2005.
22. M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, 26:148–153, 1998.
23. J. Tabaska, R. Cary, H. Gabow, and G. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–699, 1998.
24. F. van Batenburg, A. Gultyaev, C. Pleij, J. Ng, and J. Oliehoek. Pseudobase: a database with RNA pseudoknots. *Nucleic Acids Res.*, 28:201–204, 2000.
25. 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.