



**Definition 1** (RNA Secondary Structure). A secondary structure for an RNA sequence  $s$  is a set of base pairs  $\mathcal{S} = \{i \cdot j \mid 1 \leq i < j \leq |s| \wedge i < j - 3\}$ . For  $i \cdot j, i' \cdot j' \in \mathcal{S}$  with  $i \cdot j \neq i' \cdot j'$

- $\{i, j\} \cap \{i', j'\} = \emptyset$  (each base pairs with at most one other base)
- $\{s[i], s[j]\} \in \{\{A, U\}, \{C, G\}, \{G, U\}\}$  (only Watson-Crick and  $G, U$  wobble base pairs)

The second requirement, that only canonical base pairs are allowed, is standard but not consequential in solutions to the problem.

## 2.2 Scoring Schemes

Structures are usually assessed by extending the model of Gibbs free energy used for RNA Secondary Structure Prediction by Minimum Free Energy (cf. corresponding entry) with *ad hoc* extrapolation of multibranch loop energies to pseudoknot substructures [11], or by summing independent contributions e.g. obtained from base pair restricted minimum free energy structures from each base pair [13]. To investigate the complexity of pseudoknot prediction the following three simple scoring schemes will also be considered:

**Number of Base Pairs**,  $\#BP(\mathcal{S}) = |\mathcal{S}|$

**Number of Stacking Base Pairs**  $\#SBP(\mathcal{S}) = |\{i \cdot j \in \mathcal{S} \mid i + 1 \cdot j - 1 \in \mathcal{S} \vee i - 1 \cdot j + 1 \in \mathcal{S}\}|$

**Number of Base Pair Stackings**  $\#BPS(\mathcal{S}) = |\{i \cdot j \in \mathcal{S} \mid i + 1 \cdot j - 1 \in \mathcal{S}\}|$

These scoring schemes are inspired by the fact that stacked pairs are essentially the only loops having a stabilising contribution in the Gibbs free energy model.

**Problem 1** (Pseudoknot Prediction).

INPUT: RNA sequence  $s$  and an appropriately specified scoring scheme.

OUTPUT: A secondary structure  $\mathcal{S}$  for  $s$  that is optimal under the scoring scheme specified.

## 3 KEY RESULTS

**Theorem 1.** The complexities of pseudoknot prediction under the three simplified scoring schemes can be classified as follows, where  $\Sigma$  denotes the alphabet.

	Fixed alphabet	Unbounded alphabet
$\#BP$ [13]	Time $O( s ^3)$ , space $O( s ^2)$	Time $O( s ^3)$ , space $O( s ^2)$
$\#SBP$ [7]	Time $O( s ^{1+ \Sigma ^2+ \Sigma ^3})$ , space $O( s ^{ \Sigma ^2+ \Sigma ^3})$	<b>NP hard</b>
$\#BPS$	<b>NP hard</b> for $ \Sigma  = 2$ , PTAS [7] 1/3-approximation in time $O( s )$ [6]	<b>NP hard</b> [7], 1/3-approximation in time and space $O( s ^2)$ [6]

**Theorem 2.** If structures are restricted to be planar, i.e. the graph with the bases of the sequence as nodes and base pairs and backbone links of consecutive bases as edges is required to be planar, pseudoknot prediction under the  $\#BPS$  scoring scheme is **NP hard** for an alphabet of size 4. Conversely, a 1/2-approximation can be found in time  $O(|s|^3)$  and space  $O(|s|^2)$  by observing that an optimal pseudoknot free structure is a 1/2-approximation [6].

There are no steric reasons that RNA secondary structures should be planar, and the structure in Fig. 1 is actually non-planar. Nevertheless, known real structures have relatively simple overlapping base pair patterns with very few non-planar structures known. Hence, planarity has been used as a defining restriction on pseudoknotted structures [2, 15]. Similar reasoning has led to development of several algorithms for finding an optimal structure from restricted classes of structures. These algorithms tend to use more realistic scoring schemes, e.g. extensions of the Gibbs free energy model, than the three simple scoring schemes considered above.

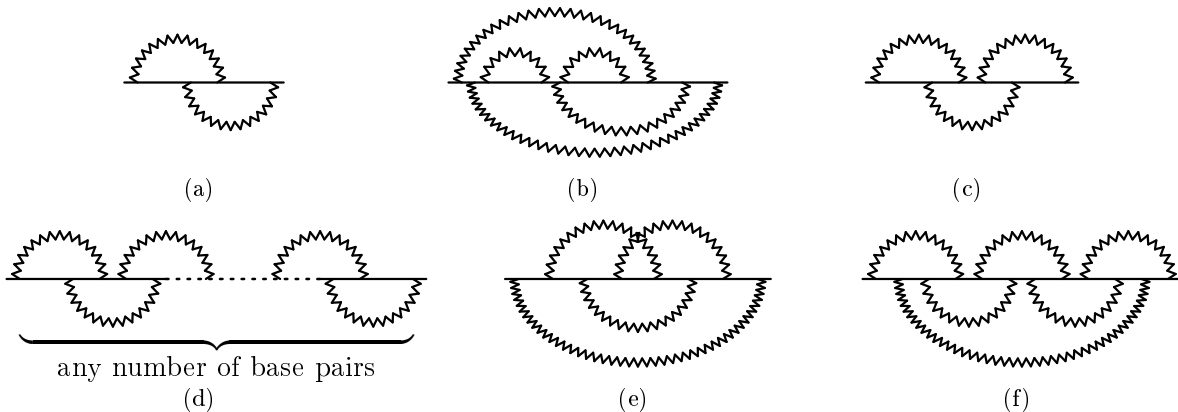


Figure 2: RNA secondary structures illustrating restrictions of pseudoknot prediction algorithms. Backbone is drawn as a straight line while base pairings are shown with zigzagged arcs.

**Theorem 3.** *Pseudoknot prediction for a restricted class of structures including Fig. 2(a) through Fig. 2(e), but not Fig. 2(f), can be done in time  $O(|s|^6)$  and space  $O(|s|^4)$  [11].*

**Theorem 4.** *Pseudoknot prediction for a restricted class of planar structures including Fig. 2(a) through Fig. 2(c), but not Fig. 2(d) through Fig. 2(f), can be done in time  $O(|s|^5)$  and space  $O(|s|^4)$  [14].*

**Theorem 5.** *Pseudoknot prediction for a restricted class of planar structures including Fig. 2(a) and Fig. 2(b), but not Fig. 2(c) through Fig. 2(f), can be done in time  $O(|s|^5)$  and space  $O(|s|^4)$  or  $O(|s|^3)$  [1, 4] (methods differ in generality of scoring schemes that can be used).*

**Theorem 6.** *Pseudoknot prediction for a restricted class of planar structures including Fig. 2(a), but not Fig. 2(b) through Fig. 2(f), can be done in time  $O(|s|^4)$  and space  $O(|s|^2)$  [1, 8].*

**Theorem 7.** *Recognition of structures belonging to the restricted classes of Theorems 3, 5, and 6, and enumeration of all irreducible cycles (i.e. loops) in such structures can be done in time  $O(|s|)$  [3, 9].*

## 4 APPLICATIONS

As for the prediction of RNA secondary structures without pseudoknots, the key application of these algorithms are for predicting the secondary structure of individual RNA molecules. Due to the steep complexities of the algorithms of Theorems 3 through 6, these are less well suited for genome scans than prediction without pseudoknots.

Enumerating all loops of a structure in linear time also allows scoring a structure in linear time, as long as the scoring scheme allows the score of a loop to be computed in time proportional to its size. This has practical applications in heuristic searches for good structures containing pseudoknots.

## 5 OPEN PROBLEMS

Efficient algorithms for prediction based on restricted classes of structures with pseudoknots that still contain a significant fraction of all known structures is an active area of research. Even using the more theoretical simple  $\#SBP$  scoring scheme, developing e.g. an  $O(|s|^{|\Sigma|})$  algorithm for this problem would be of practical significance. From a theoretical point of view, the complexity of planar structures is the least well understood, with results for only the  $\#BPS$  scoring scheme.

Classification of and realistic energy models for RNA secondary structures with pseudoknots are much less developed than for RNA secondary structures without pseudoknots. Several recent papers have been addressing this gap [3, 9, 12].

## 6 DATA SETS

PseudoBase at [biology.leidenuniv.nl/~batenburg/PKB.html](http://biology.leidenuniv.nl/~batenburg/PKB.html) is a repository of representatives of most known RNA structures with pseudoknots.

## 7 URL to CODE

The method of Theorem 3 is available at [selab.janelia.org/software.html#pknots](http://selab.janelia.org/software.html#pknots), of one of the methods of Theorem 5 at [www.nupack.org](http://www.nupack.org), and an implementation applying a slight heuristic reduction of the class of structures considered by the method of Theorem 6 is available at [bibiserv.techfak.uni-bielefeld.de/pknotsrg/](http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/) [10].

## 8 CROSS REFERENCES

RNA Secondary Structure Prediction by Minimum Free Energy.

## 9 RECOMMENDED READING

- [1] T. AKUTSU, *Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots*, Discrete Applied Mathematics, 104 (2000), pp. 45–62.
- [2] M. BROWN AND C. WILSON, *RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search*, in Biocomputing: Proceedings of the 1996 Pacific Symposium, L. Hunter and T. Klein, eds., Big Island of Hawaii, 1996, World Scientific Publishing Co, pp. 109–125.
- [3] A. CONDON, B. DAVY, B. RASTEGARI, F. TARRANT, AND S. ZHAO, *Classifying RNA pseudoknotted structures*, Theoretical Computer Science, 320 (2004), pp. 35–50.
- [4] R. M. DIRKS AND N. A. PIERCE, *A partition function algorithm for nucleic acid secondary structure including pseudoknots*, Journal of Computational Chemistry, 24 (2003), pp. 1664–1677.
- [5] T. C. GLUICK AND D. E. DRAPER, *Thermodynamics of folding a pseudoknotted mRNA fragment*, Journal of Molecular Biology, 241 (1994), pp. 246–262.
- [6] S. IEONG, M.-Y. KAO, T.-W. LAM, W.-K. SUNG, AND S.-M. YIU, *Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs*, in Proceedings of the 2nd Symposium on Bioinformatics and Bioengineering, 2001, pp. 183–190.
- [7] R. B. LYNGSØ, *Complexity of pseudoknot prediction in simple models*, in Proceedings of the 31th International Colloquium on Automata, Languages and Programming (ICALP), 2004, pp. 919–931.
- [8] R. B. LYNGSØ AND C. N. S. PEDERSEN, *RNA pseudoknot prediction in energy based models*, Journal of Computational Biology, 7 (2000), pp. 409–428.
- [9] B. RASTEGARI AND A. CONDON, *Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications*, Journal of Computational Biology, (to appear).
- [10] J. REEDER AND R. GIEGERICH, *Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics*, BMC Bioinformatics, 5 (2004), p. 104.

- [11] E. RIVAS AND S. EDDY, *A dynamic programming algorithm for RNA structure prediction including pseudoknots*, Journal of Molecular Biology, 285 (1999), pp. 2053–2068.
- [12] E. A. RØDLAND, *Pseudoknots in RNA secondary structure: Representation, enumeration, and prevalence*, Journal of Computational Biology, 13 (2006), pp. 1197–1213.
- [13] J. E. TABASKA, R. B. CARY, H. N. GABOW, AND G. D. STORMO, *An RNA folding method capable of identifying pseudoknots and base triples*, Bioinformatics, 14 (1998), pp. 691–699.
- [14] Y. UEMURA, A. HASEGAWA, S. KOBAYASHI, AND T. YOKOMORI, *Tree adjoining grammars for RNA structure prediction*, Theoretical Computer Science, 210 (1999), pp. 277–303.
- [15] C. WITWER, I. L. HOFACKER, AND P. F. STADLER, *Prediction of consensus RNA secondary structures including pseudoknots*, IEEE Transactions on Computational Biology and Bioinformatics, 1 (2004), pp. 66–77.