社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (32)
2007－BIO－11 (32)
2007／12／21

整数計画法によるシュードノットつきRNA２次構造予測

プンサップ・アンヤーニー† 加藤 有己† 阿久津 達也†

†京都大学 化学研究所 バイオインフォマティクスセンター

概要 生体高分子の機能の解明にはその折り畳み構造を理解する必要があるとされている．特に，機能的非コードRNAが注目を集めている．RNAの立体構造を予測することは困難であるため，シュードノットを含む，または含まない２次構造を予測する研究が行われてきた．本稿では，整数計画法を２次構造予測に適用する手法を提案する．ここで，シュードノットを含まない構造と，任意の平面的シュードノット構造を予測するための２つの定式化を導入する．さらに，提案手法を使った構造予測に関するいくつかの実験結果を示す．

Prediction of RNA Secondary Structure with Pseudoknots Using Integer Programming

Unyanee Poolsap†, Yuki Kato† and Tatsuya Akutsu†

†Bioinformatics Center, Institute for Chemical Research, Kyoto University

Abstract Understanding the function of biological molecules requires knowledge of their folded structures. In particular, noncoding functional RNAs have received much attention. Due to the difficulty in predicting the three dimensional structure of RNA, research efforts have shifted to the prediction of secondary structure both with and without pseudoknots. In this paper, we present a method of applying integer programming (IP) to RNA secondary structure prediction. We introduce a method with respective IP formulations for predicting pseudoknot-free structure as well as arbitrary planar pseudoknotted structure. Furthermore, we show some experimental results on structure prediction using the proposed method.

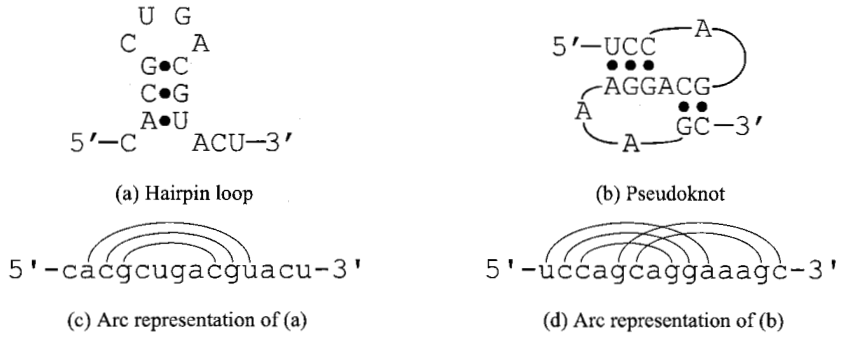
1 Introduction

One major task in bioinformatics is to predict the structure of biological molecules because the knowledge of its structure is required in order to understand how its function performs. An RNA is a biological molecule that plays many important roles in cell. The three dimensional structure of RNA is a key to understand its biological function. Unfortunately, it is very difficult to model and predict three dimensional structure; therefore, most research has been made on RNA secondary structure prediction. Analysis of secondary structure can be viewed as an intermediate step to understand its three dimensional structure, and provides information about its function.

A molecule of RNA can be viewed as a single strand of the nucleotides (bases) adenine (A), guanine (G), cytosine (C) and uracil (U). The sequence of bases is called the *primary structure*. A and U, C and G, and G and U can form a base pair via hydrogen bonding. Due to this property, the primary structure of RNA can fold back on itself to form the *secondary structure*. The secondary structure of RNA can be decomposed into stacking pairs (i.e., two or more consecutive base pairs), and a few types of loops that connect stacking pairs: hairpin loop, bulge loop, interior loop and multi-branched loop (see Figure 1 (a)). An alternative graphic representation of secondary structure is shown in Figure 1 (b), where arcs above the sequence connect base pairs. Also, there are substructures called *pseudoknots* where some base pairs occur in a crossed fashion (see Figure 1 (b), (d)).

An often-used thermodynamic hypothesis states that the actual secondary structure of an RNA sequence has the minimum free energy, where stacking pairs and loops have their associated free energy values. In general, stacking pairs have negative free energy that contributes to structure stabilization, while loop substructures have positive free energy that leads to destabilize the structure.

The problem of RNA secondary structure prediction is modeled as an energy minimization problem, and many algorithms have been developed to solve it. RNA secondary structure without pseudoknot can be predicted in $O(n^3)$ time using dynamic programming algorithms (where n is length of the sequence) [7, 8, 11]. However, it has been recognized that pseudoknots appear in many RNA molecules. Allowing pseudoknots to occur in the secondary structure causes the prediction problem harder. Several existing algorithms can predict RNA secondary structure with pseudoknots in $O(n^4)$, $O(n^5)$ or $O(n^6)$ time [1, 6, 9, 10] (also see [2]).



(a) Hairpin loop

(b) Pseudoknot

(c) Arc representation of (a)

(d) Arc representation of (b)

Figure 1: Example of RNA secondary structures

Moreover, prediction of arbitrary *planar* secondary structure including pseudoknots is proven to be NP-hard [1].

We propose a method of secondary structure prediction both with and without pseudoknots based on integer programming (IP), which is known to be NP-hard. Despite the theoretical drawbacks, it is practical and reasonable to model the prediction of planar RNA pseudoknotted structure by IP formulation.

This paper is organized as follows. First, in the preliminaries section, we formally define RNA secondary structure. Next, an IP formulation of this prediction problem is explained. The section thereafter shows the experimental results. Finally, a conclusion and future work are discussed.

2 Preliminaries

Definition 1. (RNA secondary structure)
An RNA sequence is represented by a string of n characters $s = s_1 s_2 \cdots s_n$ where $s_i \in \{A, C, G, U\}$. A *secondary structure* of a sequence s is defined as a set S of base pairs (s_i, s_j) such that the following conditions are satisfied:

1. $1 \leq i < j \leq n$, meaning, two bases that form a pair must be located at different positions.
2. $j - i > t$, where t is a small positive constant, meaning, the sequence does not fold too sharply on itself.
3. For all base pairs (s_i, s_j) and $(s_{i'}, s_{j'})$ in S, $i = i'$ if and only if $j = j'$, meaning, (s_i, s_j) and $(s_{i'}, s_{j'})$ are the same base pair.

We allow only *valid* base pairs: Watson-Crick base pairs (A, U) and (C, G), and a wobble base pair (G, U) to form in the structure.

We next define a pseudoknot, which is a kind of complex substructure of RNA.

Definition 2. (Pseudoknot)
An RNA secondary structure S is said to contain a *pseudoknot* if and only if there exist $(s_i, s_j), (s_{i'}, s_{j'}) \in S$ $(i < i')$ such that $i < i' < j < j'$. Moreover, a pseudoknotted structure is *planar* if and only if every arc can be drawn such that no two arcs cross on the plane in its arc representation[1].

On the other hand, S is called *pseudoknot-free* if and only if for all pairs $(s_i, s_j), (s_{i'}, s_{j'}) \in S$ $(i < i')$, one of the following conditions is satisfied:

1. $i < j < i' < j'$, i.e., (s_i, s_j) precedes $(s_{i'}, s_{j'})$, or
2. $i < i' < j' < j$, i.e., (s_i, s_j) includes $(s_{i'}, s_{j'})$

It is obvious that all pseudoknot-free secondary structures are planar.

[1]The pseudoknot shown in Figure 1 (d) is planar since the last two arcs can be drawn below the sequence.

	A-U	C-G	G-C	G-U	U-G	U-A
A-U	-0.9	-2.1	-1.7	-0.5	-0.9	-1.0
C-G	-1.8	-2.9	-2.0	-1.2	-1.7	-1.9
G-C	-2.3	-3.4	-2.9	-1.4	-2.1	-2.1
G-U	-1.1	-2.1	-1.9	-0.4	-1.0	-1.5
U-G	-1.1	-2.3	-1.8	-0.8	-0.9	-1.1
U-A	-0.8	-1.4	-1.2	-0.2	-0.5	-0.4

Figure 2: Energy parameter matrix E [12]

3 Integer Programming Based-Model

As mentioned in Section 1, the problem of RNA secondary structure prediction is modeled as an energy minimization problem based on a thermodynamic approach. In our integer programming-based approach, we employ the stacking energy parameters for RNA folding at 37°C as given in *Mfold* Version 2.3 [12]. We formulate two minimization integer programming (IP) models: the pseudoknot-free model and the pseudoknotted model. These two models have the same objective function, but differ in their sets of variables and constraints.

3.1 Pseudoknot-Free Model

We first introduce a set of integer variables for a mathematical formulation of the prediction problem. Let $x_{ij} = 1$ if and only if the base at position i pairs with the base at position j, otherwise $x_{ij} = 0$. In order to take stacking energy into consideration, we use $k \times l$ (row$\times$column, resp.) square energy parameter matrix E shown in Figure 2, where k and l denote the type of all possible pairs and their values range from 1 to ${}_4C_2 = 6$. For example, type 1 denotes A-U pair. We introduce z_{ij}^{kl} that corresponds to the stacking pair of (s_i, s_j) and (s_{i+1}, s_{j-1}), and $z_{ij}^{kl} = 1$ if and only if the base at position i pairs with the base at position j with type k, and the bases at the positions $i+1$ and $j-1$ pairs with type l. Let L_{x_i} and $R_{x_i} = 1$ if and only if the base at position i pairs with some base at any other position greater than i and less than i respectively. The goal is to find an assignment of 0 or 1 to all variables; although the variables of interest are $\{x_{ij}\}$. We can formulate an IP problem for pseudoknot-free structure prediction as follows:

$$\text{minimize} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{6}\sum_{l=1}^{6}(E)_{k,l}z_{ij}^{kl}$$

$$\text{subject to} \quad z_{ij}^{kl} \le \frac{x_{ij} + x_{i+1,j-1}}{2} \tag{1}$$

$$((s_i, s_j) \text{ is type } k \text{ and } (s_{i+1}, s_{j-1}) \text{ is type } l;\ i, j = 1, 2, \ldots, n;\ k, l = 1, 2, \ldots, 6),$$

$$x_{ij} = 0 \tag{2}$$

$$((s_i, s_j) \notin \{(A,U), (U,A), (C,G), (G,C), (G,U), (U,G)\};\ i, j = 1, 2, \ldots, n),$$

$$\sum_{j=1}^{i-1} x_{ji} + \sum_{j=i+1}^{n} x_{ij} \le 1 \quad (i = 1, 2, \ldots, n), \tag{3}$$

$$x_{ij} + x_{i'j'} \le 1 \quad (\forall i < i' < j < j';\ i, j, i', j' = 1, 2, \ldots, n), \tag{4}$$

$$L_{x_i} - \sum_{j=i+1}^{n} x_{ij} = 0 \quad (i = 1, 2, \ldots, n), \tag{5}$$

$$R_{x_i} - \sum_{j=1}^{i-1} x_{ji} = 0 \quad (i = 1, 2, \ldots, n), \tag{6}$$

$$L_{x_{i-1}} + (1 - L_{x_i}) + L_{x_{i+1}} > 0 \quad (i = 1, 2, \ldots, n), \tag{7}$$

$$R_{x_{i-1}} + (1 - R_{x_i}) + R_{x_{i+1}} > 0 \quad (i = 1, 2, \ldots, n), \tag{8}$$

$$x_{ij}, z_{ij}^{kl}, L_{x_i}, R_{x_i} \in \{0, 1\} \quad (i, j = 1, 2, \ldots, n;\ k, l = 1, 2, \ldots, 6). \tag{9}$$

Constraint (1) means that if base pair (s_i, s_j) is type k and (s_{i+1}, s_{j-1}) is type l, the energy parameter associated with the (k, l) stacking type will contribute to the total energy of the structure. Constraint (2) means that base pairs other than defined valid pairs cannot occur in the structure. Constraint (3) means that a base at any position in the sequence can participate in only one pair. Constraint (4) means we do not allow crossing pairs in the structure. In constraints (5) and (6), L_{x_i} and R_{x_i} are defined respectively. Constraints (7) and (8) mean that if a base at any position pairs with any other position, its previous or its next position or both must also form a base pair. Constraint (9) guarantees all variables to be either 0 or 1.

3.2 Pseudoknotted Model

For the pseudoknotted model, we add variables y_{ij} and other set of variables associated with y_{ij}, i.e., L_{y_i} and R_{y_i} to the model. The difference between x_{ij} and y_{ij} is that x_{ij} represents an arc that connects the bases above the sequence, while y_{ij} represents an arc below the sequence, which corresponds to modeling for arbitrary planar pseudoknotted structures. The objective function remains the same as the pseudoknot-free model, while the constraints are changed as follows:

$$\text{minimize} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{6}\sum_{l=1}^{6}(E)_{k,l}z_{ij}^{kl}$$

$$\text{subject to} \quad z_{ij}^{kl} \leq \frac{x_{ij} + x_{i+1,j-1}}{2}$$

$$((s_i, s_j) \text{ is type } k \text{ and } (s_{i+1}, s_{j-1}) \text{ is type } l;\ i, j = 1, 2, \ldots, n;\ k, l = 1, 2, \ldots, 6),$$

$$z_{ij}^{kl} \leq \frac{y_{i,j} + y_{i+1,j-1}}{2}$$

$$((s_i, s_j) \text{ is type } k \text{ and } (s_{i+1}, s_{j-1}) \text{ is type } l;\ i, j = 1, 2, \ldots, n;\ k, l = 1, 2, \ldots, 6),$$

$$x_{ij}, y_{ij} = 0$$

$$((s_i, s_j) \notin \{(A,U), (U,A), (C,G), (G,C), (G,U), (U,G)\};\ i, j = 1, 2, \ldots, n),$$

$$\sum_{j=1}^{i-1} x_{ji} + \sum_{j=i+1}^{n} x_{ij} + \sum_{j=1}^{i-1} y_{ji} + \sum_{j=i+1}^{n} y_{ij} \leq 1 \quad (i = 1, 2, \ldots, n),$$

$$x_{ij} + x_{i'j'} \leq 1 \quad (\forall i < i' < j < j';\ i, j, i', j' = 1, 2, \ldots, n),$$

$$y_{ij} + y_{i'j'} \leq 1 \quad (\forall i < i' < j < j';\ i, j, i', j' = 1, 2, \ldots, n),$$

$$L_{x_i} - \sum_{j=i+1}^{n} x_{ij} = 0 \quad (i = 1, 2, \ldots, n),$$

$$R_{x_i} - \sum_{j=1}^{i-1} x_{ji} = 0 \quad (i = 1, 2, \ldots, n),$$

$$L_{y_i} - \sum_{j=i+1}^{n} y_{ij} = 0 \quad (i = 1, 2, \ldots, n),$$

$$R_{y_i} - \sum_{j=1}^{i-1} y_{ji} = 0 \quad (i = 1, 2, \ldots, n),$$

$$L_{x_{i-1}} + (1 - L_{x_i}) + L_{x_{i+1}} > 0 \quad (i = 1, 2, \ldots, n),$$

$$R_{x_{i-1}} + (1 - R_{x_i}) + R_{x_{i+1}} > 0 \quad (i = 1, 2, \ldots, n),$$

$$L_{y_{i-1}} + (1 - L_{y_i}) + L_{y_{i+1}} > 0 \quad (i = 1, 2, \ldots, n),$$

$$R_{y_{i-1}} + (1 - R_{y_i}) + R_{y_{i+1}} > 0 \quad (i = 1, 2, \ldots, n),$$

$$x_{ij}, y_{ij}, z_{ij}^{kl}, L_{x_i}, L_{y_i}, R_{x_i}, R_{y_i} \in \{0, 1\} \quad (i, j = 1, 2, \ldots, n;\ k, l = 1, 2, \ldots, 6).$$

The constraint set of the pseudoknotted model has similar meaning to that of the pseudoknot-free model, except that the constraints also apply to y_{ij} variables and their associated variables (L_{y_i} and R_{y_i}).

Table 1: Prediction accuracy for the pseudoknot-free model

Seq.	Length	Avg. time (sec.)	Sensitivity (%)	Specificity (%)	F-measure (%)
RF00032	26	0.157	100	100	100
RF00037	30	0.296	60	60	60
RF00196	35	0.508	93	95	94
RF00385	42	0.989	97	99	98
RF00250	56	3.667	97	98	98

Table 2: Prediction accuracy for the pseudoknotted model

Seq.	Length	Time (sec.)	Sensitivity (%)	Specificity (%)	F-measure (%)
BaMV	39	3.080	25.00	27.27	26.09
BMV3_UPD_PK1	26	0.346	80.00	88.89	84.21
BMV3_UPD_PK2	21	0.174	25.00	50.00	33.33
BMV3_UPD_PK3	23	0.259	0	0	0
BMV3_UPD_PK4	26	0.453	88.89	72.73	80.00
BSBV2_UPD_PKb	33	0.998	40.00	44.44	42.11
BSBV2_UPD_PKc	24	0.292	66.67	66.67	66.67
BSBV3_UPD_PKb	36	9.400	83.33	66.67	74.07
BSBV3_UPD_PKc	24	0.288	77.78	100.00	87.50
BSMVbeta_UPD_PK1	32	1.492	0	0	0
BSMVbeta_UPD_PK2	26	0.408	100.00	100.00	100.00
BSMVbeta_UPD_PK3	33	1.511	30.00	42.86	35.29
STMV_UPD1_PK1	25	0.346	71.43	50.00	58.82
STMV_UPD1_PK2	26	0.422	77.78	58.33	66.67
STMV_UPD1_PK3	32	0.977	33.33	42.86	37.50
STMV_UPD2_PK1	31	1.065	0	0	0
STMV_UPD2_PK2	26	0.597	0	0	0
STMV_UPD2_PK3	24	0.262	100.00	80.00	88.89
STNV1_PK1	26	0.371	0	0	0
STNV1_PK2	31	0.832	75.00	69.23	72.00
STNV1_PK3	26	0.417	20.00	28.57	23.53
STNV2_PK1	29	0.500	0	0	0
STNV2_PK2	25	0.301	88.89	80.00	84.21
STNV2_PK3	27	0.425	33.33	60.00	42.86
TMGMV_UPD_PK1	27	0.378	50.00	55.56	52.63
TMGMV_UPD_PK2	24	0.260	77.78	77.78	77.78
TMGMV_UPD_PK3	33	1.148	37.50	33.33	35.29

4 Experimental Results

Both models described in Section 3 were tested on sets of sequences with known structure. The pseudoknot-free sequences were obtained from Rfam [5]. We selected 5 families of short RNA sequences. For each family, 5 sequences were chosen randomly. On the other hand, sequences that are known to contain pseudoknots were obtained from PseudoBase [3]. 29 short sequences were selected randomly. We used the ILOG CPLEX 10.1 [13] software to solve the IP models.

We evaluate the prediction results by calculating sensitivity, specificity and F-measure. Sensitivity is the number of correctly predicted base pairs divided by the number of base pairs of the known structure. Specificity is the number of correctly predicted base pairs divided by the total number of predicted base pairs. F-measure is the harmonic mean of sensitivity α and specificity β, which is defined as $\frac{2\alpha\beta}{\alpha+\beta}$.

Table 1 and Table 2 shows the results of the pseudoknot-free model and the pseudoknotted model respectively. According to the sensitivity and specificity values, the pseudoknot-free model yields good prediction results for all of the test sequence sets. For the pseudoknotted model, although some structures were predicted correctly or almost correctly, a few structures cannot be predicted.

5 Conclusion

We introduced two integer programming (IP) models for RNA secondary structure prediction: the pseudoknot-free model and the pseudoknotted model. We performed tests on our prediction models with sets of known structure sequences, which were selected randomly from the databases. We then used the sensitivity and specificity to evaluate the predicted results. Since the performance of the pseudoknotted model is not as high as expected, we have to reconsider the model. In addition, we have to include other kinds of energy parameters into both models in order to improve the prediction accuracy.

References

[1] Akutsu, T.: Dynamic Programming Algorithms for RNA Secondary Structure Prediction with Pseudoknots, *Discrete Applied Mathematics*, Vol. 104, pp. 45–62 (2000).

[2] Akutsu, T.: Recent Advances in RNA Secondary Structure Prediction with Pseudoknots, *Current Bioinformatics*, Vol. 1, No. 2, pp. 115–129 (2006).

[3] Batenburg, F. H. D. van, Gultyaev, A. P., Pleij, C. W. A, Ng, J. and Oliehoek, J.: Pseudobase: A Database with RNA Pseudoknots, *Nucleic Acids Research*, Vol. 28, No. 1, pp. 201–204 (2000).

[4] Clote, P. and Backofen, R.: *Computational Molecular Biology: An Introduction*, Wiley West Sussex, UK (2000).

[5] Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. and Bateman, A.: Rfam: Annotating Non-Coding RNAs in Complete Genomes, *Nucleic Acids Research*, Vol. 33, pp. D121–124 (2005).

[6] Kato, Y., Seki, H. and Kasami, T.: RNA Pseudoknotted Structure Prediction Using Stochastic Multiple Context-Free Grammar, *IPSJ Transactions on Bioinformatics*, Vol. 47, No. SIG17 (TBIO1), pp. 12–21 (2006).

[7] Mount, D. W.: *Bioinformatics: Sequence and Genome Analysis,* Cold Spring Harbor Laboratory Press, New York, USA (2001).

[8] Nussinov, R., Pieczenik, G., Griggs, J. R. and Kleitman, D. J.: Algorithms for Loop Matchings, *SIAM Journal of Applied Mathematics*, Vol. 35, No. 1, pp. 68–82 (1978).

[9] Rivas, E. and Eddy, S. R.: A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots, *Journal of Molecular Biology*, Vol. 285, pp. 2053–2068 (1999).

[10] Uemura, Y., Hasegawa, A., Kobayashi, S. and Yokomori, T.: Tree Adjoining Grammars for RNA Structure Prediction, *Theoretical Computer Science*, Vol. 210, pp. 277–303 (1999).

[11] Zuker, M. and Stiegler, P.: Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information, *Nucleic Acids Research*, Vol. 9, pp. 133–148 (1981).

[12] http://frontend.bioinfo.rpi.edu/zukerm/cgi-bin/efiles.cgi

[13] http://www.ilog.com/products/cplex