IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming

Kengo Sato1,*,†, Yuki Kato2,*,†, Michiaki Hamada1, Tatsuya Akutsu2 and Kiyoshi Asai1,4

1Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, 2Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikaruna, Nara 630-0192, 3Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, and 4Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan

ABSTRACT

Motivation: Pseudoknots found in secondary structures of a number of functional RNAs play various roles in biological processes. Recent methods for predicting RNA secondary structures cover certain classes of pseudoknotted structures, but only a few of them achieve satisfying predictions in terms of both speed and accuracy.

Results: We propose IPknot, a novel computational method for modeling a wide class of pseudoknots based on maximizing expected accuracy of a predicted structure. IPknot decomposes a pseudoknotted structure into a set of pseudoknot-free substructures and approximates a base-pairing probability distribution that considers pseudoknots, leading to the capability of modeling a wide class of pseudoknots and running quite fast. In addition, we propose a heuristic algorithm for refining base-pairing probabilities to improve the prediction accuracy of IPknot. The problem of maximizing expected accuracy is solved by using integer programming with threshold cut. We also extend IPknot so that it can predict the consensus secondary structure with pseudoknots when a multiple sequence alignment is given. IPknot is validated through extensive experiments on various datasets, showing that IPknot achieves better prediction accuracy and faster running time as compared with several competitive prediction methods.

Availability: The program of IPknot is available at http://rna.naist.jp/ipknot/. IPknot is also available as a web server at http://rna.naist.jp/ipknot/. Contact: sato@info.u-tokyo.ac.jp; ykato@iis.naist.jp

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

RNAs play various roles in biological processes, ranging from the passive role as a messenger that carries genetic information to the active roles as a regulator for gene expression and as a catalyst in cellular processes. Considerable attention has been paid to the functions of RNAs, especially those of regulatory non-coding RNAs (Eddy, 2001). It is widely believed that there is a strong correlation between the 3D structure of an RNA molecule and its function. A set of base pairs formed from hydrogen bonds is called a secondary structure, which shapes the substructure of the 3D structure. Since experimental determination of RNA 3D structures is difficult and their structures are hierarchical, secondary structure prediction provides a major key to elucidating the potential functions of RNAs.

A good number of computational studies have so far been presented on RNA secondary structure prediction. They can be roughly classified into two groups, namely comparative sequence analysis and single sequence analysis. Comparative methods based on alignment folding include RNAalifold (Bernhart et al., 2008; Hofacker et al., 2002) and Pfold (Knudsen and Hein, 2003). The comparative analysis has an advantage of a fair possibility of achieving high prediction accuracy since it can take evolutionary information into consideration. However, this approach is not always applicable since a set of homologous sequences is required in advance. As for single sequence analysis, a popular approach is to find the structure with the minimum free energy (MFE) of a single RNA sequence. We can use programs that implement this strategy based on dynamic programming (DP) algorithms such as mfold (Zuker, 2003; Zuker and Stiegler, 1981) and RNAfold (Hofacker, 2003; Hofacker et al., 1994). Notice that the free energy of a secondary structure is calculated by summing energy parameters of respective loop substructures, which can be experimentally determined and computationally estimated (Mathews et al., 1999). Furthermore, these DP-based methods are applied to the calculation of the partition function for RNA secondary structures (McCaskill, 1990), which enables us to compute posterior base-pairing probabilities. Recently, several sophisticated methods have been proposed for predicting the secondary structure with the maximum expected accuracy (MEA) over a space of possible structures. CONTRAfold (Do et al., 2006) and CentroidFold (Hamada et al., 2009a) that adopt this idea achieve better prediction accuracy as compared with the MFE-based methods. It is to be noted that all of the above methods aim to predict relatively simple RNA structures with nested base-pairing interactions.

Pseudoknot is one of the important topologies in RNA secondary structures. A pseudoknot is typically formed from the base pairings between the unpaired bases of a loop and those outside the loop, which is often called an H-type pseudoknot (see Fig. 1a). In other words, a secondary structure includes a pseudoknot if at least two arcs drawn above the primary sequence that represent base pairs cross each other (see Fig. 1b). Pseudoknotted structures are observed in many RNAs such as ribosomal RNAs, transfer messenger RNAs and viral RNAs (van Batenburg et al., 2001). Pseudoknots are known
Designing RNA energy models is also an important task to predict secondary structures of good quality. An energy model consists of structural features (fragments), energy parameters and a function that assigns a free energy change to a structure of a given sequence (Andronescu et al., 2010b). The Mathews–Turner model (Mathews et al., 1999) is widely used to predict RNA secondary structures without pseudoknots. The Dirks–Pierce model (Dirks and Pierce, 2003) includes the Mathews–Turner features and additional features for pseudoknots. The Cao–Chen model (Cao and Chen, 2006) includes the Dirks–Pierce features along with many new features for H-type pseudoknots. Recently, Andronescu et al. (2007, 2010b) have presented algorithms for refining energy parameters using a constraint generation approach and Boltzmann likelihood estimation. Moreover, Andronescu et al. (2010a) have reported that HotKnots employing new energy parameters estimated by these training algorithms yields better prediction accuracy on pseudoknotted structural data as compared with the earlier version of HotKnots.

Although there are various approaches to predicting RNA pseudoknotted structures, only a few of them achieve satisfying predictions in both speed and accuracy. This is a crucial requirement especially when applying a prediction algorithm to finding functional non-coding RNAs in genome sequences. In this article, we present IPknot, a novel method for Integer Programming (IP)-based prediction of RNA pseudoKNOTs. IP, one of the optimization techniques, is useful for modeling a wide variety of combinatorial problems. Remember that one of the existing methods mentioned above uses IP to predict RNA pseudoknotted structures (Pool sap et al., 2009). Main differences between the earlier IP-based study and our newly proposed method lie in the objective functions and the classes of pseudoknots that they can handle. The important point to note is that IPknot significantly outperforms the earlier IP-based method in both prediction accuracy and running time. As in our previous method HotKnots (Kato et al., 2010) for RNA–RNA interaction prediction, IPknot seeks to find the MEA secondary structure using IP. To compute the expected accuracy of a secondary structure with respect to an ensemble of all possible structures including pseudoknots, we decompose a pseudoknotted structure into a set of pseudoknot-free substructures and approximate a base-pairing probability distribution that considers pseudoknots. This decomposition enables IPknot to describe a wide class of pseudoknotted structures and perform quite fast predictions.

To find the consensus secondary structure for pseudoknots when a multiple alignment of RNA sequences is given, employing the methodology of CentroidAlifold (Hamada et al., 2011) for pseudoknot-free consensus structure prediction. We validate the prediction performance of IPknot through extensive experiments on various datasets, making a comparison with several state-of-the-art prediction methods. The major advantages of this work in performance are summarized as follows:

- Prediction performance of IPknot is sufficiently good in speed and accuracy as compared with ProbKnot (Bellaousov and Mathews, 2010), FlexStem (Chen et al., 2008),
After discussing the results in Section 4, we conclude this article in Section 5. We present a new method IPknot for predicting pseudoknotted RNA secondary structures using integer programming (IP). IPknot executes the following two steps when an RNA sequence is given:

(i) compute the base-pairing probabilities used in the IP objective function (Section 2.1);
(ii) solve the IP problem to predict the optimal pseudoknotted RNA secondary structure (Section 2.2).

In Section 2.3, we propose a heuristic algorithm for refining the base-pairing probabilities that compose the IP objective function in the first step. Furthermore, we extend our algorithm to common secondary structure prediction including pseudoknots in Section 2.4.

2 METHODS

We present a new method IPknot for predicting pseudoknotted RNA secondary structures using integer programming (IP). IPknot executes the following two steps when an RNA sequence is given:

(i) compute the base-pairing probabilities used in the IP objective function (Section 2.1);
(ii) solve the IP problem to predict the optimal pseudoknotted RNA secondary structure (Section 2.2).

In Section 2.3, we propose a heuristic algorithm for refining the base-pairing probabilities that compose the IP objective function in the first step. Furthermore, we extend our algorithm to common secondary structure prediction including pseudoknots in Section 2.4.

2.1 MEA-based scoring function for predicting pseudoknotted RNA secondary structures

Let $\Sigma = \{A,C,G,U\}$ and $\Sigma^*$ denote the set of all finite RNA sequences consisting of bases in $\Sigma$. For a sequence $x = x_1x_2\ldots x_n \in \Sigma^*$, let $|x|$ denote the number of symbols appearing in $x$, which is called the length of $x$. Let $\mathcal{S}(x)$ be a set of secondary structures of an RNA sequence $x$ including pseudoknots. An element $y \in \mathcal{S}(x)$ is represented as a $|x| \times |x|$ binary-valued triangular matrix $y = (y_{ij})$, where $y_{ij} = 1$ means that bases $x_i$ and $x_j$ form a base pair.

We assume that a secondary structure $y \in \mathcal{S}(x)$ can be decomposed into a set of pseudoknot-free substructures $(y^{(1)}, y^{(2)}, \ldots, y^{(m)})$ that satisfies the following conditions: (i) $y \in \mathcal{S}(x)$ should be decomposed into a mutually-exclusive set, that is, for all $1 \leq i \neq j \leq |S|$, $\sum_{y = 1}^{m} y^{(i)}$ and $\sum_{y = 1}^{m} y^{(j)} \leq 1$; and (ii) every base pair in $y^{(i)}$ should be pseudoknotted to at least one base pair in $y^{(j)}$ for $i < j$. Each pseudoknot-free substructure $y^{(i)}$ is said to belong to the level $p$ (see Fig. 2). For any RNA secondary structure $y \in \mathcal{S}(x)$, there exists a positive integer $m$ such that $y$ can be decomposed into $m$ pseudoknot-free substructures [see Supplementary Section S6 and Jiang et al. (2010) for further details]. From this viewpoint, we can assert that the above decomposition enables our method to model arbitrary pseudoknots.

One of the most promising techniques to predict RNA secondary structures is the MEA-based approach including centroid estimation (Carvalho and Lawrence, 2008; Hamada et al., 2009a).

First, we define a gain function $G_{y}(y_{ij})$ of the secondary structure $y \in \mathcal{S}(x)$ with regard to the correct secondary structure $y \in \mathcal{S}(x)$ as follows:

$$G_{y}(y_{ij}) = y^{T}P(y_{ij}) + TN(y_{ij})$$

where $y = 0$ is a weight parameter for base pairs, $TP$ and $TN$ denote the numbers of true positives (base pairs) and true negatives (non-base pairs), respectively, and $I(condition)$ is the indicator function that takes a value of 1 or 0 depending on whether the condition is true or false.

Our objective is to find a secondary structure $\hat{y}$ that maximizes the expected gain function (1) under a given probability distribution over the space $\mathcal{S}(x)$ of pseudoknotted secondary structures:

$$\hat{y} = \text{arg max}_{y \in \mathcal{S}(x)} \sum_{y \in \mathcal{S}(x)} G_{y}(y_{ij})P(y_{ij})$$

where $P(y_{ij})$ is a probability distribution of RNA secondary structures including pseudoknots. It has been proven that the $\gamma$-centroid estimator (2) enables us to decode accurate secondary structures from a given probability distribution (Hamada et al., 2009a).

Unfortunately, the calculation of Equation (2) is intractable for arbitrary pseudoknots (Akufo, 2000; Lyngsø and Pedersen, 2000b). Instead, we can employ several models for limited classes of pseudoknots such as the Rivas–Eddy model (Rivas and Eddy, 1999), the Akutsu model (Akutsu, 2000), the Dirks–Pierce model (Dirks and Pierce, 2003, 2004) and the Reeder–Giegerich model (Reeder and Giegerich, 2004). However, even for relatively simple pseudoknotted structures, computationally expensive costs of $O(|x|^3)$ time and $O(|x|^2)$ space are required. Therefore, we approximate the expected gain function (2) by the sum of the expected gain functions for each level of pseudoknot-free substructures $(y^{(1)}, \ldots, y^{(m)})$ in the decomposed set of a pseudoknotted structure $y \in \mathcal{S}(x)$, and thus simultaneously find a pseudoknotted structure $\hat{y}$ and its decomposition $(y^{(1)}, \ldots, y^{(m)})$ that maximize:

$$\sum_{y = 1}^{m} G_{y}(y_{ij})P(y_{ij})$$

where $y^{(i)} > 0 \sum_{y = 1}^{m} y^{(i)} = 1$ is a weight parameter for each gain function at the level $p$ (in our experiments, we fixed $y^{(i)} = 1/m$, $y^{(i)} > 0$ is a weight parameter for base pairs at the level $p$, and $C$ is a constant independent of $\gamma$ [see the Supplementary Material of Hamada et al. (2009a) for the derivation]).

The base-pairing probability $P(y_{ij})$ is a probability that the base $i$ is paired with $j$, which is defined as:

$$p_{ij} = \sum_{y \in \mathcal{S}(x)} I(j,y_{ij})P(y_{ij})$$

We can select $P(y_{ij})$, a probability distribution over a set $\mathcal{S}(x)$ of secondary structures with or without pseudoknots, from several approaches. A naive
approach is of use of the probability distribution with pseudoknots as well as a pseudoknot (2) in spite of high computational costs. Alternatively, we can employ a probability distribution without pseudoknots such as the McCaskill model (McCaskill, 1990), whose computational complexity is $O(|\Sigma|^3)$ for time and $O(|\Sigma|^2)$ for space. Furthermore, a novel approach that refines the base-pairing probability matrix from the distribution without pseudoknots will be proposed in Section 2.3. Note that we implemented all the three approaches.

It is worth mentioning that IPknot can be regarded as an extension of CentroidFold (Hamada et al., 2009a). If we let the number of decomposed levels $m = 1$, the approximate expected gain function (3) is identical to the $\gamma$-centroid estimator used in CentroidFold.

We should notice that the maximization of the approximate gain (3) is equivalent to the maximization of the weighted sum of the base-pairing probabilities $p_i$ larger than $\theta^0 = 1/\sum_{i=1}^{\ell} q_i + 1$. Consequently, it is no longer necessary to consider the base pairs whose pairing probabilities are at most the threshold $\theta^0$, which we call threshold cut.

### 2.2 IP model

Maximization of the approximate expected gain (3) can be solved by the IP problem as follows:

\[
\max \sum_{1 \leq i \leq n} \alpha_i p_i^0 \sum_{x \in S^a} \prod_{i=1}^{\ell} \gamma_{ij}^{0}\left( \sum_{x \in S^a} \prod_{i=1}^{\ell} \gamma_{ij}^{0} \right) \leq 1 \quad (1 \leq y \leq n),
\]

\[
\frac{\gamma_{ij}^{0}}{\gamma_{ij}^{0}} \leq 1 \\
(1 \leq i \leq n, 1 \leq j \leq m).
\]

Since Equation (4) is an instantiation of the approximate estimator (3) and the threshold cut technique is applicable to Equation (3), we need to consider only base pairs $y_{ij}^0$ whose base-pairing probabilities $p_i$ are larger than $\theta^0 = 1/\sum_{i=1}^{\ell} q_i + 1$. The constraint (5) means that each base $x_i$ can be paired with at most one base (Fig. 3a). The constraint (6) disallows pseudoknots within the same level $p$ (Fig. 3b). The constraint (7) ensures that each base pair at the level $p$ is pseudoknotted to at least one base pair at every lower level $q \prec q'$. (Fig. 3c).

It is widely accepted that base pairs in stable RNA structures are likely to appear in a stacked form rather than an isolated one. Following the IP formulation proposed by Poolesh et al. (2009), we can avoid isolated base pairs by incorporating the stacked pairing constraints as follows:

\[
\ell_{ij}^0 + 1 - \ell_{ij}^0 + \ell_{ij}^0 + 1 = 1 \\
(1 \leq i < n),
\]

\[
\ell_{ij}^0 + 1 - \ell_{ij}^0 + \ell_{ij}^0 + 1 = 1 \\
(1 < i < n),
\]

These constraints guarantee that if a base $x_i$ is paired with another one, the base(s) adjacent to $x_i$ must also form a base pair.

### 2.3 An iterative refinement algorithm for the base-pairing probability matrix

We propose an iterative algorithm that refines the base-pairing probabilities used in the objective function of our method. The basic idea is that the base-pairing probabilities are improved by the secondary structures predicted at the first stage, and then a new prediction is performed by the improved base-pairing probabilities (see also Fig. 4).

More specifically, for a given sequence $x \in \Sigma^*$, we first predict a secondary structure $\gamma = (\gamma_1, \ldots, \gamma_\ell)$ by solving the IP problem described as Equations (4)-(9). Then, for each level $p$, a constraint $\theta^0$ on secondary structure is constructed as follows: (i) for $i, j$ such that $\gamma_{ij}^0 = 1$, only the base pair between $x_i$ and $x_j$ is allowed, and other base pairs that involve $x_i$ or $x_j$ are disallowed; and (ii) for $i, j$ such that $\gamma_{ij}^0 = 1$ ($q \neq p$), $x_i$ and $x_j$ cannot form base pairs at the level $p$. The base-pairing probabilities with the constraint $\theta^0$ can be defined as $p_i^0 = \sum_{x \in S^a} \prod_{i=1}^{\ell} \gamma_{ij}^{0}$ (see also Fig. 4). This calculation is performed by a variant of the McCaskill algorithm in $O(|\Sigma|^3)$ time and $O(|\Sigma|^2)$ space. Finally, using the updated base-pairing probabilities $p_i = \sum_{x \in S^a} \prod_{i=1}^{\ell} \gamma_{ij}^{0}$, we re-predict a secondary structure $\gamma$ by solving the IP problem. These steps are iterated until an eligible condition (e.g. the number of iterations, the convergence of the prediction) is satisfied.

The probability distribution produced by the iterative refinement algorithm can be regarded as a mixture of the probability distribution for each level $p$ of the preceding prediction of pseudoknot-free structures, which can represent a wider space of the distribution than the individual distribution of pseudoknot-free structures. This enables the iterative refinement algorithm to improve the base-pairing probability matrix.

### 2.4 Common secondary structure prediction including pseudoknots

It is well known that use of multiple alignments of homologous sequences improves the accuracy of predicting RNA secondary structures due to the alignment information such as covariation (Bernhart et al., 2008; Hamada et al., 2011). In order to implement the prediction of common secondary structures including pseudoknots for aligned sequences, we can apply the
We validated IPknot. The first term of Equation (10) contributes to the robustness against the alignment errors, and the second term improves the probability distribution by the covariance information on each column of the alignment.

We predict the optimal common secondary structure including pseudoknots for aligned RNA sequences by solving the IP problem introduced in Section 2.2 with $p_{ij}^X$ instead of $p_{ij}$. Note that the iterative refinement algorithm described in Section 2.3 can work as well as the case of individual sequences.

3 RESULTS

3.1 Implementation

Our method was implemented as a program called IPknot. The first dataset was obtained from the RNA STRAND database (Andronescu et al., 2008), which contains only carefully curated sequences and structures. We selected the RNA sequences with at least one pseudoknot whose length is between 140 nt and 500 nt. To reduce redundant sequences, we filtered out the sequences that have more than 85% identity to the other sequences. As a result, 388 non-redundant sequences were obtained. We call this ‘RS-pk388’ dataset.

The second dataset is called ‘pk168’ dataset, originally established by Huang and Ali (2007). The pk168 dataset is compiled from PseudoBase (van Batenburg et al., 2001), which includes 16 categories of pseudoknots. After excluding the redundant sequences (> 85% identity), the test set includes 168 sequences whose lengths are <140 nt. This dataset was also used by recent studies (Chen et al., 2008; Huang and Ali, 2007).

For the benchmark of prediction of common secondary structures including pseudoknots for aligned RNA sequences, we created the third dataset from Rfam 10.0 (Gardner et al., 2011). Only manually curated seed alignments with consensus structures published in literature were used. We produced 67 alignments containing five sequences from the Rfam families that satisfy the following conditions: (i) at least one pseudoknot is included; (ii) the length is at most 500 nt; and (iii) at least five sequences are contained. We call this ‘Rfam-RPK’ dataset. In order to evaluate the robustness against the alignment errors, we realigned every alignment by ClustalW (Thompson et al., 1994), which considers no structural information such as covariation.

3.3 Prediction of secondary structures including pseudoknots

The experiment on the RS-pk388 dataset was conducted, comparing our algorithm with several state-of-the-art methods that can predict pseudoknots including PseudoBase (Bellaousov and Mathews, 2010), FlexStem (Chen et al., 2008), HotKnots (Gardner et al., 2008) and ILM (Ruan et al., 2006), and those that can predict only pseudoknot-free structures including CentroidFold (Hamada et al., 2009a) and RNAfold (Hofacker, 2003).

For IPknot, we fixed the number of decomposed sets of secondary substructures $m = 2$, and varied the weight parameters for the expected number of true positive base pairs in such a way that $\gamma = [2^{k}] |k = 0, 1, 2, 3, 4]$. Since CentroidFold has the weight parameters for the expected number of true positive base pairs as well as IPknot, the same range of parameters was applied to these two methods. For HotKnots, IPknot9 parameters estimated by Andronescu et al. (2010a) were employed. For the other competitive methods, the default settings were used.

We evaluated prediction accuracy through positive predictive value (PPV) and sensitivity (Sen) with respect to base pairs defined as follows:

$$\text{PPV} = \frac{TP}{TP + FP}, \quad \text{Sen} = \frac{TP}{TP + FN}$$

where $TP$ is the number of correctly predicted base pairs, $FP$ is the number of incorrectly predicted base pairs, and $FN$ is the number of base pairs in the true structure that were not predicted.

Figure 5 shows the PPV-Sensitivity plots for respective algorithms. Note that the sets of points with the same shape plotted for IPknot and CentroidFold correspond to the results obtained by changing values of the weight parameters $\gamma$. The results clearly indicate that IPknot is more accurate than the existing methods on the RS-pk388 dataset. It can also be seen that the iterative refinement algorithm improves the prediction accuracy of IPknot.

Since the pk168 dataset contains shorter sequences than the RS-pk388 dataset, more accurate but computationally expensive
prediction algorithms can be applied to this dataset. For the pk168 dataset, we evaluated IPknot with the D&P model (Dirks and Pierce, 2003, 2004) that calculates the exact base-pairing probabilities including a limited class of pseudoknots. Furthermore, our method was compared with the DP-based algorithms including NUPACK (Dirks and Pierce, 2003, 2004) and PHOTOS (Rivas and Eddy, 1999), as well as the competitive methods used in the previous experiment. Figure 6 shows the accuracy on the pk168 dataset, indicating that IPknot with the D&P model is comparable to pknotsRG, which achieves the best accuracy.

We evaluated the time efficiency of IPknot using GLPK as the IP solver and the competitive methods on Linux OS with Intel Quad Core Xeon E5450 (3.0 GHz) and 32 GB memory. Five sequences PKB229, PKB134 (from the pk168 dataset), ASE_00193, CFK_00614 (from the RS-pk388 dataset) and CFX_00774 (from the RNA STRAND database) were used to measure the elapsed time to predict secondary structures. Table 1 indicates that IPknot is significantly faster than the existing algorithms for predicting pseudoknotted secondary structures. It should be noted that the reason why IPknot with the D&P model takes long running time is not due to solving the IP problem but due to computing the exact base-pairing probabilities that consider pseudoknots.

3.4 Prediction of common secondary structures including pseudoknots

A few number of algorithms for common secondary structure prediction with pseudoknots have been available. In this experiment, we compared IPknot with hxmatch (Witwer et al., 2004) and ILM (Ruan et al., 2004) for pseudoknotted common secondary structure prediction in addition to CentroidAlifold (Hamada et al., 2011) and RNAalifold (Bernhart et al., 2008) for pseudoknot-free common secondary structure prediction. We evaluated the accuracy through PPV and sensitivity for common secondary structures by mapping them to the individual sequences in the multiple alignments. The experimental results are shown in Figure 7. For the hand-curated reference alignments, hxmatch and IPknot with the iterative refinement algorithm achieve almost the same level of accuracy (see Fig. 7a). However, the alignments of low quality produced by ClustalW cause significantly worse accuracy of hxmatch compared with IPknot (Fig. 7b). Furthermore, in the Supplementary paper, the results of the experiments on multiple alignments produced by ProbCons (Do et al., 2005) optimized for non-coding RNAs (called ProbConsRNA) and CentroidAlign (Hamada et al., 2009b) for RNA structural alignments are shown in Supplementary Figure S3. These results suggest that IPknot has the robustness against the alignment errors compared with the existing methods such as hxmatch and ILM.

4 DISCUSSION

IPknot predicts a pseudoknotted secondary structure that maximizes the approximate expected gain function (3), which represents the expectation of the (weighted) number of true predictions of base pairs under a given probability distribution. We can regard this approach as maximizing expected accuracy, which has been successfully applied into various problems (Carvalho and Lawrence, 2008; Do et al., 2005; Do et al., 2006, 2008; Hamada et al., 2009a, b, 2010, 2011; Kato et al., 2010; Knudsen and Hein, 2003). Recent studies have revealed that MEA-based methods can achieve more accurate predictions than other methods such as the maximum a posterior (MAP)-based and MFE-based methods, even from the same probability distribution. In fact, as shown in Figure 6, IPknot with the D&P model is much superior to NUPACK, both of which employ the same probability distribution (the D&P model) but different in decoding algorithms (based on MEA and MFE, respectively).

The threshold cut technique that enables IPknot to run fast is derived from Equation (3), which suggests that too large $\gamma^{AP}$ is not suitable for the balanced accuracy measures such as Matthews Correlation Coefficient (MCC) and F-measure. It should
Table 1. Comparison of time performance between different algorithms

<table>
<thead>
<tr>
<th>ID</th>
<th>PEB229</th>
<th>PEB134</th>
<th>ASE_00193</th>
<th>CRW_00614</th>
<th>CRW_00774</th>
</tr>
</thead>
<tbody>
<tr>
<td>length (nt)</td>
<td>67</td>
<td>137</td>
<td>301</td>
<td>494</td>
<td>989</td>
</tr>
<tr>
<td>IPknot (without refinement)</td>
<td>0.01 s</td>
<td>0.04 s</td>
<td>0.19 s</td>
<td>0.63 s</td>
<td>6.40 s</td>
</tr>
<tr>
<td>IPknot (with refinement)</td>
<td>0.01 s</td>
<td>0.05 s</td>
<td>0.28 s</td>
<td>0.94 s</td>
<td>18.0 s</td>
</tr>
<tr>
<td>(with D&amp;P model)</td>
<td>8.63 s</td>
<td>8 m 26 s</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>ProbKnot</td>
<td>0.06 s</td>
<td>0.32 s</td>
<td>4.52 s</td>
<td>11 m 30 s</td>
<td>29.3 s</td>
</tr>
<tr>
<td>FlexStem</td>
<td>0.49 s</td>
<td>0.68 s</td>
<td>5.24 s</td>
<td>10 m 17 s</td>
<td>15 m 28 s</td>
</tr>
<tr>
<td>HotKnots</td>
<td>4.24 s</td>
<td>44.0 s</td>
<td>125 m 12 s</td>
<td>125 m 10 s</td>
<td>N/A</td>
</tr>
<tr>
<td>pknotsARG</td>
<td>0.02 s</td>
<td>0.28 s</td>
<td>3.29 s</td>
<td>24 s</td>
<td>6 m 50 s</td>
</tr>
<tr>
<td>ILM</td>
<td>0.02 s</td>
<td>0.12 s</td>
<td>0.21 s</td>
<td>1.32 s</td>
<td>23.9 s</td>
</tr>
<tr>
<td>NUPACK</td>
<td>1.91 s</td>
<td>24.1 s</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>CentroidFold</td>
<td>0.01 s</td>
<td>0.04 s</td>
<td>0.19 s</td>
<td>0.59 s</td>
<td>6.36 s</td>
</tr>
<tr>
<td>RNAfold</td>
<td>&lt;0.01 s</td>
<td>0.01 s</td>
<td>0.07 s</td>
<td>0.21 s</td>
<td>0.85 s</td>
</tr>
</tbody>
</table>

N/A means that we were unable to complete calculation on our machine.

Fig. 7. The PPV–Sensitivity plots of the experiment on the Rfam-PK dataset. (a) The results for reference alignments. (b) The results for sequence-based alignments produced by Clustal W.

be emphasized that base pairs whose probability is less than the threshold will make the accuracy degenerate in expectation. Using the threshold cut technique, IPknot as well as RactIP (Kato et al., 2010) makes the IP problem so sparse that practical problems such as RNA pseudoknot secondary structure prediction and RNA–RNA joint secondary structure prediction can be solved even by the IP solver that is freely available but inferior in performance.

Comparative experiments with the existing methods on single RNA sequences show that IPknot produces better predictions in accuracy on long RNA sequences, whereas its relative accuracy drops on short sequences, especially for IPknot without refinement. One possible explanation for these results is that for short sequences, many base pairs with high probabilities are likely to be predicted at the level 1, resulting in scarcity of base pairs predicted at the level 2. Nevertheless, the iterative refinement algorithm enhances the base-pairing probabilities at the level 2 by masking the base pairs predicted at the level 1, leading to improve prediction accuracy on short sequences.

As shown in Section 3.4, IPknot can perform robust predictions of consensus structures when a multiple alignment of RNA sequences is given. This is mainly due to use of Equation (10) that represents the mixture distribution of sequence-based and alignment-based probabilities. Unlike the competitive methods based only on covariance information, the averaged base-pairing probabilities described in the first term of Equation (10) play an important role in keeping the quality of prediction when the alignment quality gets worse.
IPknot has several parameters that users should select, including the weights for true positive base pairs at the level $p^{(m)}$ in Equation (3), the number of decomposed levels of pseudoknots, and the number of iterations of the iterative refinement algorithm. As well as CentroidFold and CentroidAllFold, the weights $y^{(m)}$ for true positive base pairs control the balance of PPV and sensitivity. Since a set of larger $y^{(m)}$ predicts more base pairs, sensitivity will increase. However, false positive base pairs will also increase and thus will make PPV decrease. Furthermore, the threshold values of the base-pairing probabilities to consider in the object function (4) depend on $y^{(m)}$. This means that if we select too large $y^{(m)}$, the performance of IPknot will degenerate since the IP problem to solve enlarges. Therefore, it is a crucial issue to determine appropriate parameters $y^{(m)}$, although we showed the results on several sets of $y^{(m)}$ in this article. To this end, we can take two approaches. First, use of the best pseudo-expected MCC (Hamada et al., 2010) among several sets of $y^{(m)}$ can be mentioned. The pseudo-expected MCC is a good approximation of expected MCC, which can be calculated from the base-pairing probabilities. As the second approach, we can use machine learning techniques such as the max-margin method (Do et al., 2008), which will adopt the parameters to given training datasets.

As shown in Section 3, the improved base-pairing probability matrices by the iterative refinement algorithm made the prediction accuracy much elevated, especially for short sequences. In these experiments, relatively small $y^{(m)}$ such as a pair of $y^{(1)} = 1$ and $y^{(2)} = 1$ achieved favorable performance, suggesting that use of only probable base pairs predicted by large thresholds at the first step would produce reliable base-pairing probabilities for the second step. Another important factor in the performance of IPknot is the number of iterations of the iterative refinement algorithm. Significant improvement was observed when applying the refinement algorithm once as compared with no refinement. On the other hand, we could not find meaningful difference between running the algorithm once and twice as shown in Supplementary Figure S4. Since each iteration is a time-consuming procedure, it seems reasonable to suppose that performing the iteration algorithm once is sufficient.

The maximum complexity of a secondary structure predicted by IPknot is restricted by the number $m$ of decomposed levels of pseudoknots, which is also called an $m$-partite RNA secondary structure (Jiang et al., 2010), defined as the union of $m$ pseudoknot-free substructures. A recent study has implied that most known RNA secondary structures are either bipartite or tripartite, i.e. $m$-partite for $m = 2$ or $3$ (Redland, 2006). Supplementary Figure S5 in the supplementary paper shows the experiments on the RS-pkfrel41 (pseudoknot-free) dataset and the RS-pk388 (pseudoknoted) dataset for the maximum decomposed level $m = 1,2,3$. Note again that CentroidFold is equivalent to IPknot with the maximum decomposed level $m = 1$. It can be observed that the selection of a conflicting level between predicted structures and correct structures causes the degradation of the accuracy compared with the best results. For example, for the pseudoknot-free dataset, IPknot with the maximum decomposed level $m = 3$ and the iterative refinement algorithm cut down the accuracy compared with CentroidFold because of increasing false positive base pairs. These results indicate that the appropriate number of decomposed levels should be selected, although correct structures might be unknown. Redland (2006) has revealed that among the hundreds of known RNA secondary structures with pseudoknots in Pseudobase (van Batenburg et al., 2001), only one structure is tripartite and others are all bipartite. This fact suggests that IPknot will work satisfactorily on average if we select the number of decomposed levels $m = 2$.

## 5 CONCLUSION

We proposed a new computational method IPknot for predicting RNA secondary structures with a wide class of pseudoknots, which can take either a single sequence or aligned sequences as input. We demonstrated using a variety of structural datasets that IPknot is sufficiently fast and accurate as a computational prediction tool for both single sequence analysis and comparative sequence analysis. Prediction accuracy of IPknot depends mainly on its scoring functions even though the method uses the approximate probability distribution for pseudoknoted structures. In fact, experimental results revealed that IPknot with the base-pairing probabilities computed by heuristic refinement produces much better predictions than that without refinement. Moreover, when we adopted the exact probabilities for pseudoknots, a significant improvement in accuracy was confirmed, though much computation time was spent on the predictions. Considering these results, there is room for further investigation into refinement of the scoring functions that make prediction accuracy compatible with running time.

Another important fact to stress is that IPknot can run quite fast even on a relatively long sequence less than one thousand bases. This is attributed to use of both approximation of a probability distribution for pseudoknots and integer programming with threshold cut based on maximizing expected accuracy. As described in Section 1, prediction methods with satisfying speed and accuracy are useful for finding functional non-coding RNA genes from genome sequences. Making skillful use of the speed and accuracy of our method, exhaustive search for genes of non-coding RNAs that may form pseudoknots is a worthwhile task as another future work.

## ACKNOWLEDGEMENT

We wish to thank our colleagues from the RNA Informatics Team at the Computational Biology Research Center (CBRC) for fruitful discussions. The super-computing resource was provided by Human Genome Center, Institute of Medical Science, University of Tokyo.

**Funding:** Grant-in-Aid for Young Scientists (B) (KAKENII) from Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan; Global COE program ‘Deciphering Biosphere from Genome Big Bang’ from MEXT, Japan; Grant-in-Aid for Scientific Research on Innovative Areas from MEXT, Japan (to M.H. and K.A., in part).

**Conflict of Interest:** none declared.

## REFERENCES


Andronescu M. et al. (2010a) Improved free energy parameters for RNA pseudoknotted secondary structure prediction. RNA, 16, 26–42.