Vol.2010-BIO-21 No.32
2010/6/19

RNA-RNA Interaction Prediction Using Integer Programming with Threshold Cut

Yuki Kato*1,†1 Kengo Sato*1,†2
Michiaki Hamada,†3,†4 Yoshihide Watanabe,†5
Kiyoshi Asai†2,†4 and Tatsuya Akutsu†1

Much attention has been focused on predicting RNA-RNA interaction since it is a key to identifying possible targets of noncoding small RNAs that regulate gene expression post-transcriptionally. A number of computational studies have so far been devoted to predicting joint secondary structures or binding sites under a specific class of interactions. In this technical report, we propose `RactIP`, a fast and accurate prediction method for RNA-RNA interaction of general type based on integer programming. `RactIP` can integrate approximate information on an ensemble of equilibrium joint structures into the objective function using posterior internal and external base paring probabilities. Experimental results show that prediction accuracy of `RactIP` is at least comparable to that of several state-of-the-art methods for RNA-RNA interaction prediction. Moreover, we demonstrate that `RactIP` can run incomparably faster than competitive methods for predicting joint secondary structures.

1. Introduction

In recent years there has been a renewal of interest in the biological roles of functional noncoding RNAs (ncRNAs). A new light on their functions thrown by modern studies has given us recognition that they can act as ubiquitous regulators in living cells[12),24)]. A class of small ncRNAs downregulates gene expression post-transcriptionally via base pairing with its target mRNAs to inhibit the translation into the corresponding proteins. Several regulatory *antisense RNAs* have been found in bacterial chromosomes, which have relatively long sequences and interact with their target mRNAs in a more intricate manner[6)]. In particular, *kissing hairpin* structures (see Fig. 1) caused by loop-loop interaction have been observed[7)]. To help to understand the mechanism of RNA-RNA interaction further as well as to identify target RNAs of specific ncRNAs, it is desirable to develop fast and accurate computational methods for predicting interacting RNA structures.

RNA-RNA interaction prediction has so far been performed by several computational approaches, which can be roughly classified into four groups. The first group belongs to the category that uses the idea of concatenating two RNA sequences and considering them as a single strand so that the minimum free energy (MFE) structure of the resulting sequence can be computed. `PairFold`[3)] and `RNAcofold`[5)] adopt this procedure, but these methods cannot predict general type of interaction such as kissing hairpins. The second group including `UNAFold`[10)], `RNAhybrid`[22)] and `RNAduplex` from the Vienna RNA Package[13),14)] disregards intramolecular bonds in both two sequences and computes the MFE hybridization. They work out for short interacting RNAs but are impracticable for long sequences that have intramolecular structures. Other approaches such as `RNAup`[18)], `IntaRNA`[8)], `inRNAs`[23)] and `bistaRNA`[21)] fall into the third group, which considers RNA-RNA interaction as the stepwise process of individual intramolecular foldings and their hybridization. The final group aims at predicting the MFE joint secondary structure or computing the interaction partition function under the comprehensive type of interaction. `IRIS`[19)], `inteRNA`[2)], `RIG`[17)], `piRNA`[9)], `rip`[15),16)] and also `inRNAs`[23)] are listed as this category. These methods impose natural constraints on the joint structure such that there are no internal pseudoknots, crossing interactions and zigzags (see 2), 9) for details). Note that Alkan *et al.* suggested that these constraints are satisfied by all examples of complex RNA-RNA interactions in the literature. In this sense, we can consider the class of joint structures satisfying those constraints the most general type of interaction. Although the final group can cover wider class of interacting structures simultaneously, their computational costs could be prohibitively expensive

†1 Bioinformatics Center, Institute for Chemical Research, Kyoto University
†2 Graduate School of Frontier Sciences, University of Tokyo
†3 Mizuho Information & Research Institute, Inc
†4 Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST)
†5 Department of Mathematical Sciences, Faculty of Science and Engineering, Doshisha University
∗1 The authors wish it to be known that in their opinion the first two authors should be regarded as joint First Authors.

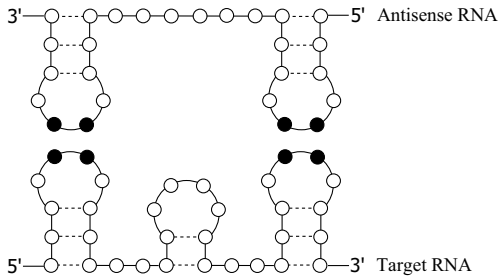

Fig. 1 An example of RNA-RNA interaction containing two kissing hairpins cited from 21). A broken line represents an internal base pair, and a black circle indicates a base that constitutes an external base pair (binding site).

for long sequences.

Prediction of interacting RNA structure can be considered as a kind of optimization problem in a sense that we seek to minimize the free energy of the joint structure or maximize a score such as an interaction probability under the possible topology of the interacting structure. Recently, Bauer *et al.*[4] presented `LARA` that implements an *integer programming*-based method for multiple structural alignment of RNA sequences. To reduce the computational cost of the structural alignment, they simplified the mathematical structure of the integer programming (IP) formulation, eliminating information on the topology of secondary structure from the constraints and integrating the structural information into the objective function, as well as using the Lagrangian relaxation.

Motivated in part by Bauer *et al.*'s work, we present a novel method called `RactIP`, RNA-RNA interACTion prediction using Integer Programming. In our IP formulation, the objective function, which is to be maximized, is defined as the sum of scores with respect to internal and external base pairs of two interacting RNAs. In particular, we make use of posterior probabilities such as *base pairing probabilities* and *hybridization probabilities* for scores of internal pairs and those of external pairs, respectively, to incorporate further structural information into the objective function as in the case of `LARA`[4]. It is to be noted that use of such posterior probabilities enables the method to take account of an ensemble of equilibrium structures approximately, which will lead to improve expected accuracy. We introduce a *threshold cut* technique to solve the IP problem efficiently, which is shown to be reasonable from the viewpoint of maximizing expected accuracy. By virtue of this technique, `RactIP` achieves considerably short run-time despite the computational hardness of IP. We demonstrate the prediction performance of `RactIP` on a set of known interacting RNA pairs both for joint structure prediction and binding site prediction, which shows that accuracy of `RactIP` is at least comparable to that of several state-of-the-art methods. Moreover, we demonstrate that `RactIP` can run overwhelmingly faster than competitive methods for predicting joint secondary structures.

2. Methods

We propose a new method `RactIP` for RNA-RNA interaction prediction based on integer programming (IP). `RactIP` executes the following two steps when two RNA sequences are given:

(1) Compute the score matrices of the IP objective function for internal and external base pairs;
(2) Solve the IP problem to predict the optimal joint secondary structure.

2.1 Scoring Functions for Predicting RNA-RNA Interaction

Let $\Sigma = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$ and Σ^* denote the set of all finite RNA sequences consisting of bases in Σ. For a sequence $a \in \Sigma^*$, let $|a|$ denote the number of symbols appearing in a, which is called the length of a. For $1 \leq i < j \leq |a|$, we let $a[i..j]$ denote a sequence $a_i a_{i+1} \cdots a_j \in \Sigma^*$.

Let $\mathcal{S}(a)$ be a space of all possible secondary structures of a sequence $a \in \Sigma^*$. An element $x \in \mathcal{S}(a)$ is represented as a $|a|^2$ binary-valued triangular matrix $x = (x_{ij})_{i<j}$, where $x_{ij} = 1$ means that bases a_i and a_j form a base pair. We denote by $P(x \mid a)$ a probability distribution over $\mathcal{S}(a)$. Let $\mathcal{H}(a, b)$ be a space of all possible hybridized structures between $a, b \in \Sigma^*$, which considers no secondary structures of a and b. An element $z \in \mathcal{H}(a, b)$ is represented as a $|a| \times |b|$ binary-valued matrix $z = (z_{ij})$, where $z_{ij} = 1$ means that the base a_i interacts with the base b_j. We denote by $P(z \mid a, b)$ a probability distribution over $\mathcal{H}(a, b)$. Let $\mathcal{JS}(a, b)$ be a space of the joint secondary structures of a and b that considers both the secondary structures of a and b and the hybridized structures between a and b. In other words, $\mathcal{JS}(a, b)$ is a subset of $\mathcal{S}(a) \times \mathcal{S}(b) \times \mathcal{H}(a, b)$. We denote by $P(\sigma \mid a, b)$ a probability distribution over $\mathcal{JS}(a, b)$ where $\sigma = (x, y, z) \in \mathcal{JS}(a, b)$

such that $x \in \mathcal{S}(a)$, $y \in \mathcal{S}(b)$ and $z \in \mathcal{H}(a,b)$. We assume that each base can be paired with at most one base regardless of whether the base pair is formed inside or outside the sequence, and internal pseudoknots and crossing interactions (external pseudoknots) are disallowed.

We now define the problem of predicting RNA-RNA interaction as follows:

Definition 1. *Given two RNA sequences $a = a[1..n] \in \Sigma^*$ (5′–3′ direction) and $b = b[1..m] \in \Sigma^*$ (3′–5′ direction), predict a joint secondary structure $\sigma \in \mathcal{JS}(a,b)$.*

To tackle this problem, we first define the gain function for the true joint structure $\sigma = (x,y,z)$ and a predicted joint structure $\hat{\sigma} = (\hat{x},\hat{y},\hat{z})$ as

$$G(\sigma,\hat{\sigma}) = G_{\gamma_s}(x,\hat{x}) + G_{\gamma_s}(y,\hat{y}) + \alpha \cdot G_{\gamma_h}(z,\hat{z}) \tag{1}$$

where

$$G_{\gamma_s}(x,\hat{x}) = \sum_{i<j} \left[\gamma_s \cdot I(x_{ij}=1)I(\hat{x}_{ij}=1) + I(x_{ij}=0)I(\hat{x}_{ij}=0)\right],$$
$$G_{\gamma_s}(y,\hat{y}) = \sum_{i<j} \left[\gamma_s \cdot I(y_{ij}=1)I(\hat{y}_{ij}=1) + I(y_{ij}=0)I(\hat{y}_{ij}=0)\right],$$
$$G_{\gamma_h}(z,\hat{z}) = \sum_{i,j} \left[\gamma_h \cdot I(z_{ij}=1)I(\hat{z}_{ij}=1) + I(z_{ij}=0)I(\hat{z}_{ij}=0)\right].$$

Here, $I(condition)$ is the indicator function that takes a value of 1 or 0 depending on whether the *condition* is true or false, γ_s and γ_h are weight parameters for base pairs, and α is a balancing parameter between internal base pairs and external base pairs. The gain function (1) is equal to the weighted sum of the number of true positives and the number of true negatives of base pairs. In order to maximize the *expected* number of true predictions, we find a joint secondary structure $\hat{\sigma}$ that maximizes the expectation of the gain function (1) with respect to an ensemble of all possible joint secondary structures under a given posterior distribution:

$$\mathbb{E}_{\sigma|a,b}[G(\sigma,\hat{\sigma})] = \sum_{\sigma \in \mathcal{JS}(a,b)} G(\sigma,\hat{\sigma})\, P(\sigma \mid a,b). \tag{2}$$

For the posterior distribution $P(\sigma \mid a,b)$ over a space of joint secondary structures, we can employ the `piRNA` model[9] and the `rip` model[15),16]. However, these exact models are impractical since $O(n^6)$ time and $O(n^4)$ space are required where n is the length of the longer RNA sequence. Therefore, we approximate the posterior distribution over a space of joint secondary structures by its factorization:

$$P(\sigma \mid a,b) \simeq P(x \mid a)\, P(y \mid b)\, P(z \mid a,b) \tag{3}$$

for $\sigma = (x,y,z)$. As a result, the expected gain (2) can be replaced by

$$\sum_{x \in \mathcal{S}(a)} G_{\gamma_s}(x,\hat{x})\, P(x \mid a) + \sum_{y \in \mathcal{S}(b)} G_{\gamma_s}(y,\hat{y})\, P(y \mid b) + \alpha \sum_{z \in \mathcal{H}(a,b)} G_{\gamma_h}(z,\hat{z})\, P(z \mid a,b)$$
$$= \sum_{i<j} \left[(\gamma_s+1)p_{ij}^{(a)} - 1\right] \hat{x}_{ij} + \sum_{i<j} \left[(\gamma_s+1)p_{ij}^{(b)} - 1\right] \hat{y}_{ij} + \alpha \sum_{i,j} \left[(\gamma_h+1)q_{ij} - 1\right] \hat{z}_{ij} + C \tag{4}$$

where C is a constant independent of $\hat{\sigma}$, $p_{ij}^{(a)} = \sum_{x \in \mathcal{S}(a)} I(x_{ij}=1)\, P(x \mid a)$ is a *base pairing probability* that the base a_i is paired with a_j, and $q_{ij} = \sum_{z \in \mathcal{H}(a,b)} I(z_{ij}=1)\, P(z \mid a,b)$ is a *hybridization probability* that the base a_i interacts with the base b_j. Finally, we find a joint structure $\hat{\sigma} = (\hat{x},\hat{y},\hat{z})$ that maximizes the approximate estimator (4). We should notice that maximizing the approximate estimator (4) is equivalent to maximizing the sum of: (i) the sum of the base pairing probabilities $p_{ij}^{(a)}$ larger than $\theta_s = 1/(\gamma_s+1)$; (ii) the sum of the base pairing probabilities $p_{ij}^{(b)}$ larger than θ_s; and (iii) the sum of the hybridization probabilities q_{ij} larger than $\theta_h = 1/(\gamma_h+1)$. This means that there is no need to consider the base pairs whose posterior probabilities are at most the predefined thresholds θ_s and θ_h so that the search space for the optimal structure can be reduced. This threshold cut technique makes our method run much faster.

2.2 Integer Programming Model

We use the entries of the binary-valued triangular matrices $x = (x_{ij})_{i<j} \in \mathcal{S}(a)$, $y = (y_{ij})_{i<j} \in \mathcal{S}(b)$ and $z = (z_{ij}) \in \mathcal{H}(a,b)$ defined in Section 2.1 as the fundamental IP variables for describing internal and external base pairs. With these variables, we can formulate an IP problem for joint secondary structure prediction as follows:

$$\text{maximize} \quad \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_{ij}^{(a)} x_{ij} + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} p_{ij}^{(b)} y_{ij} + \alpha \sum_{i=1}^{n}\sum_{j=1}^{m} q_{ij} z_{ij} \tag{5}$$

$$\text{subject to} \quad \sum_{j=1}^{i-1} x_{ji} + \sum_{k=i+1}^{n} x_{ik} + \sum_{l=1}^{m} z_{il} \le 1 \quad (1 \le \forall i \le n), \tag{6}$$

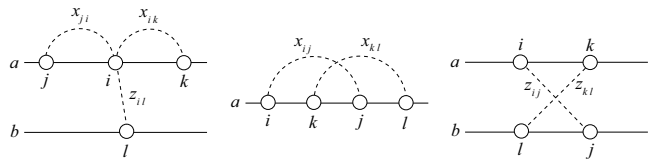

Fig. 2 An illustration of several constraints of the IP formulation. In each of the three diagrams, at most one variable shown by a broken (curved) line can take a value 1.

$$\sum_{j=1}^{i-1} y_{ji} + \sum_{k=i+1}^{m} y_{ik} + \sum_{l=1}^{n} z_{li} \leq 1 \quad (1 \leq \forall i \leq m), \tag{7}$$

$$x_{ij} + x_{kl} \leq 1 \quad (1 \leq \forall i < \forall k < \forall j < \forall l \leq n), \tag{8}$$

$$y_{ij} + y_{kl} \leq 1 \quad (1 \leq \forall i < \forall k < \forall j < \forall l \leq m), \tag{9}$$

$$z_{ij} + z_{kl} \leq 1 \quad (1 \leq \forall i < \forall k \leq n;\ 1 \leq \forall l < \forall j \leq m) \tag{10}$$

where $p_{ij}^{(a)}$, $p_{ij}^{(b)}$ and q_{ij} denote the base pairing probabilities and the hybridization probability defined in Section 2.1, respectively, and $\alpha \in (0, 1)$ is a parameter that regulates the proportion of hybridization in the predicted structure. Recall that values of all variables must be either 0 or 1. Figure 2 illustrates several constraints of the IP formulation.

To solve the IP problem, we employ the threshold cut technique where we exclude the IP variables in advance representing internal and external base pairs whose posterior probabilities are not exceeding θ_s and θ_h, respectively, defined in Section 2.1, before computing the optimal solution. As described in Section 2.1, this threshold cut is derived from the viewpoint of maximizing expected accuracy of joint structure prediction.

It is widely accepted that base pairs in stable RNA structures are likely to appear in a stacked form rather than an isolated one. We employ further variables and related constraints for promoting stacking base pairs, following the IP formulation for predicting secondary structure of a single RNA sequence proposed by Poolsap *et al.*[20].

3. Results

3.1 Implementation

Our method was implemented as a program called `RactIP`, which uses Gurobi optimizer 2.0 (`http://gurobi.com/`) for solving the integer programming problem. We employed the `CONTRAfold` model[11] for the probability distribution of RNA secondary structures and the `RNAduplex` model for the probability distribution of hybridization of two RNA sequences. `CONTRAfold`, based on a machine learning algorithm, is one of the most accurate programs for predicting RNA secondary structures. We utilized part of `CONTRAfold` to calculate base pairing probabilities $p_{ij}^{(a)}$ and $p_{ij}^{(b)}$. `RNAduplex` is a program from the Vienna RNA package[13),14)] for computing the MFE structure of hybridization of two RNA sequences. We modified the code of `RNAduplex` to calculate hybridization probabilities q_{ij} instead of the MFE structures. The source code of `RactIP` is freely available at `http://www.ncrna.org/software/ractip/`.

3.2 Joint Structure Prediction

We first conducted experiments in joint secondary structure prediction on the dataset compiled by Kato *et al.*[17]. The performance was evaluated by sensitivity and positive predictive value (PPV) defined as follows:

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

where TP is the number of correctly predicted base pairs, FN is the number of base pairs in the true structure that were not predicted, and FP is the number of incorrectly predicted base pairs. We also used F-measure as the balanced measure between sensitivity and PPV, which is defined as the harmonic mean of them:

$$F = \frac{2 \times \text{sensitivity} \times \text{PPV}}{\text{sensitivity} + \text{PPV}}.$$

We compared our method `RactIP` with two state-of-the-art methods: `inRNAs` (the exact model for joint structure prediction)[23)] and `inteRNA`[2)]. The accuracy of `inRNAs` is extracted from their literature. In order to calculate the accuracy of `inteRNA`, we computed the joint structure of each pair in the dataset by using the `inteRNA` Web server with default settings[1)] (`http://compbio.cs.sfu.ca/taverna/interna/`).

Table 1 Comparison with competitive methods for joint structure prediction. The five RNA-RNA interaction pairs were predicted by `RactIP`, `inRNAs` (the joint structure prediction model)[23] and `inteRNA` (the Loop Model)[1),2)]. We set the parameters for `RactIP` as $\alpha = 0.5$, $\theta_s = 0.5$ and $\theta_h = 0.2$. Running time of `RactIP` was measured on Mac OS X 10.6 running Intel Core 2 Duo 2.13 GHz with 2 GB memory. Note that computation time of `inRNAs` is reported to be at most 4000 seconds for long sequences on Sun Fire X4600 2.6 GHz with 64 GB memory[23].

Antisense-target	Sensitivity			PPV			F-measure			Time (s)
	RactIP	inRNAs	inteRNA	**RactIP**	inRNAs	inteRNA	**RactIP**	inRNAs	inteRNA	**RactIP**
CopA-CopT	1.000	1.000	0.731	0.754	0.846	0.655	0.860	0.917	0.691	0.13
DIS-DIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.05
IncRNA_{54}-RepZ	0.813	0.875	0.958	0.736	0.792	0.836	0.772	0.831	0.893	0.10
R1inv-R2inv	1.000	0.900	0.800	1.000	0.900	0.889	1.000	0.900	0.842	0.03
Tar-Tar*	1.000	1.000	1.000	0.875	0.875	0.875	0.933	0.933	0.933	0.03
Average	0.950	0.955	0.898	0.846	0.883	0.851	0.913	0.916	0.872	

Table 1 shows the results of joint structure prediction using our approach and two existing methods. As can be seen, `RactIP` outperforms `inteRNA` and is comparable to `inRNAs`. We did not compare the running time strictly between these methods due to difficulty in their availability. However, we would like to remark that Salari *et al.*[23] reported in their literature that `inRNAs` runs for $\sim$ 4000 seconds on Sun Fire X4600 2.6 GHz with 64 GB memory to predict the joint structures of CopA-CopT and IncRNA_{54}-RepZ. Meanwhile, `RactIP` consumes only 0.13 seconds and 0.10 seconds to predict the joint structures of CopA-CopT and IncRNA_{54}-RepZ, respectively, on Mac OS X 10.6 running on Intel Core 2 Duo 2.13 GHz with 2 GB memory.

3.3 Binding Site Prediction

In the second experiment, we assessed the performance of predicting binding sites on the dataset reported by Kato *et al.*[17] and Busch *et al.*[8]. The accuracy was measured by sensitivity, PPV and F-measure such that only external base pairs are considered. Table 2 shows the results of prediction by our program `RactIP`, `inRNAs` (the heuristic model for binding site prediction)[23] and `IntaRNA`[8], indicating that our method is more accurate or comparable as compared with the existing methods. It is worth noting that `RactIP` has no restriction on the number of accessible regions to predict, whereas `IntaRNA` and `inRNAs` can consider only one or two accessible regions that are putative binding sites.

3.4 Time and Accuracy Trade-Off by Approximation

To confirm the effectiveness of approximating the joint posterior distribution by its factorization, we compared running time and prediction accuracy of the factorized model that we proposed with those of the naïve model by `rip`[15),16)]. `rip` calculates exact base pairing probabilities of internal base pairs and external base pairs by taking $O(n^6)$ for time and $O(n^4)$ for space where n is the length of the longer sequence. We compared `RactIP` with `rip`, which samples joint structures from internal and external base pairing probabilities, and `RactIP` combined with `rip`, in which internal and external base pairing probabilities calculated by `rip` were used in the integer programming (Eq. (5)) instead of factorized ones. Note that `rip` failed to calculate base pairing probabilities for the CopA-CopT pair since the length of them might be too long for `rip`. As shown in Table 3, our approximation by factorization is significantly faster than the naïve calculation of base pairing probabilities, though the accuracy of our approximation dropped slightly. The results indicate that our method can be applicable to joint secondary structure prediction for long sequences.

4. Conclusion

We proposed `RactIP`, a novel method for predicting RNA-RNA interaction of general type using integer programming (IP). Our mathematical model is so compact that we can easily incorporate elaborate structural information into the objective function, utilizing posterior probabilities such as internal base pairing probabilities and hybridization probabilities. Furthermore, we introduced the threshold cut technique used in solving the IP problem, which greatly reduces the

Table 2 Comparison with competitive methods for RNA-RNA binding site prediction. The 23 RNA-RNA interaction pairs were predicted by `RactIP`, `inRNAs` (the binding site prediction model)[23] and `IntaRNA`[8]. We set the parameters for `RactIP` as $\alpha = 0.5$, $\theta_s = 0.3$ and $\theta_h = 0.5$. Running time of `RactIP` and `IntaRNA` was measured on Mac OS X 10.6 running on Intel Core 2 Duo 2.13 GHz with 2 GB memory. Computation time of `inRNAs` measured on Intel Core 2 Duo 2.53 GHz with 4 GB memory was given by R. Salari (personal communication).

Antisense-target	Sensitivity			PPV			F-measure			Time (s)		
	RactIP	inRNAs	IntaRNA	**RactIP**	inRNAs	IntaRNA	**RactIP**	inRNAs	IntaRNA	**RactIP**	inRNAs	IntaRNA
CopA-CopT	0.815	0.889	1.000	0.579	0.828	0.391	0.677	0.857	0.562	0.14	0.21	0.14
DIS-DIS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.04	0.03	0.04
$IncRNA_{54}$-RepZ	0.750	1.000	0.738	0.783	0.889	0.850	0.766	0.941	0.790	0.10	2.56	0.11
R1inv-R2inv	1.000	1.000	1.000	1.000	0.778	1.000	1.000	0.875	1.000	0.03	0.03	0.02
Tar-Tar*	1.000	1.000	1.000	0.833	0.833	0.833	0.909	0.909	0.909	0.02	0.03	0.03
DsrA-RpoS	0.654	0.808	0.808	0.739	0.778	0.776	0.694	0.793	0.793	0.06	6.80	0.05
GcvB-argT	0.950	0.950	0.950	0.950	0.864	0.950	0.950	0.905	0.950	0.04	8.07	0.03
GcvB-dppA	0.941	1.000	1.000	0.593	0.850	0.586	0.727	0.919	0.739	0.05	5.59	0.04
GcvB-gltI	1.000	0.750	0.000	0.828	0.500	0.000	0.906	0.600	0.000	0.05	2.74	0.04
GcvB-livJ	0.955	0.634	0.955	0.955	0.824	0.955	0.955	0.717	0.955	0.04	6.10	0.04
GcvB-livK	0.958	0.540	0.542	0.958	0.570	0.565	0.958	0.555	0.553	0.04	3.24	0.03
GcvB-oppA	1.000	1.000	1.000	1.000	0.733	0.957	1.000	0.846	0.978	0.05	8.23	0.04
GcvB-STM4351	0.880	0.760	0.760	0.957	1.000	0.905	0.917	0.864	0.826	0.04	2.59	0.04
IstR-tisAB	0.778	0.722	0.879	1.000	1.000	0.960	0.875	0.839	0.918	0.05	8.24	0.05
MicA-ompA	0.875	1.000	1.000	0.737	1.000	1.000	0.800	1.000	1.000	0.32	3.29	0.04
MicA-lamB	0.565	1.000	1.000	0.867	1.000	0.821	0.684	1.000	0.902	0.08	5.38	0.08
MicC-ompC	0.727	1.000	1.000	0.889	1.000	0.537	0.800	1.000	0.699	0.07	8.11	0.06
MicF-ompF	0.833	0.960	0.960	0.690	0.960	0.960	0.755	0.960	0.960	0.73	17.82	0.83
OxyS-fhlA	0.563	0.813	0.500	0.818	1.000	1.000	0.667	0.897	0.667	0.32	0.21	0.39
RyhB-sdhD	0.882	0.618	0.588	0.833	0.955	1.000	0.857	0.750	0.741	0.07	7.74	0.06
RyhB-sodB	1.000	1.000	1.000	0.310	1.000	0.818	0.474	1.000	0.900	0.19	3.23	0.21
SgrS-ptsG	0.739	0.566	0.739	0.895	0.765	1.000	0.810	0.651	0.850	0.06	12.07	0.05
Spot42-galK	0.682	0.432	0.409	0.698	0.760	0.643	0.690	0.551	0.500	0.13	5.94	0.13
Average	0.847	0.845	0.776	0.851	0.865	0.805	0.836	0.845	0.791			

Table 3 Comparison of accuracy and running time for joint structure prediction. The four RNA-RNA interaction pairs were predicted by `RactIP`, `rip`[15),16)] and `RactIP` with base pairing probabilities calculated by `rip` (`rip`+`RactIP`). We set the parameters for `RactIP` as $\alpha = 0.5$, $\theta_s = 0.5$ and $\theta_h = 0.2$. Running time of `RactIP` was measured on Mac OS X 10.6 running on Intel Core 2 Duo 2.13 GHz with 2 GB memory. Computation time of `rip` was measured on linux kernel 2.6.30 running on Intel Xeon 3.33 GHz with 32 GB memory.

Antisense-target	Sensitivity			PPV			F-measure			Time	
	RactIP	rip	rip+RactIP	**RactIP**	rip	rip+RactIP	**RactIP**	rip	rip+RactIP	**RactIP**	rip
DIS-DIS	1.000	0.500	0.500	1.000	0.500	0.500	1.000	0.500	0.500	0.05s	19m40s
$IncRNA_{54}$-RepZ	0.813	0.562	1.000	0.736	0.500	0.889	0.772	0.529	0.941	0.10s	860m
R1inv-R2inv	1.000	0.900	1.000	1.000	0.900	1.000	1.000	0.900	1.000	0.03s	37s
Tar-Tar*	1.000	1.000	1.000	0.875	0.875	0.875	0.933	0.933	0.933	0.03s	9.5s

complexity of the solution space. Experimental results demonstrated that prediction accuracy of `RactIP` is at least comparable to that of several state-of-the-art methods for joint structure prediction and binding site prediction. Experimental validations revealed that `RactIP` can run much faster than competitive methods for predicting joint secondary structures. `RactIP` not only achieved success in RNA-RNA interaction prediction but also showed further possibility of applying the fast IP-based method with the threshold cut technique to other biologically important problems, which are worthwhile and challenging tasks.

Acknowledgments This work was supported in part by Grant-in-Aid for Young Scientists (B) (KAKENHI) from Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan [#22700313 to Y.K.; #22700305 to K.S.]; and Grant-in-Aid for Scientific Research (A) (KAKENHI) from Japan Society for the Promotion of Science (JSPS) [#22240009 to T.A.]. The authors would like to thank Ms. Unyanee Poolsap for her assistance in collecting the experimental data. The authors also wish to thank Ms. Raheleh Salari for providing them with experimental results using the method presented by her group.

References

1) Aksay,C., Salari,R., Karakoc,E., Alkan,C. and Sahinalp,S.C. (2007) taveRNA: a web suite for RNA algorithms and applications. *Nucleic Acids Res.*, **35**, W325–W329.
2) Alkan,C., Karakoc,E., Nadeau,J.H., Sahinalp,S.C. and Zhang,K. (2006) RNA-RNA interaction prediction and antisense RNA target search. *J. Comput. Biol.*, **13**, 267–282.
3) Andronescu,M., Zhang,Z.C. and Condon,A. (2005) Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.*, **345**, 987–1001.
4) Bauer,M., Klau,G.W. and Reinert,K. (2007) Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinform.*, **8**, 271.
5) Bernhart,S.H., Tafer,H., Mückstein,U., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3.
6) Brantl,S. (2002) Antisense-RNA regulation and RNA interference. *Biochim. Biophys. Acta*, **1575**, 15–25.
7) Brunel,C., Marquet,R., Romby,P. and Ehresmann,C. (2002) RNA loop-loop interactions as dynamic functional motifs. *Biochimie*, **84**, 925–944.
8) Busch,A., Richter,A.S. and Backofen,R. (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.
9) Chitsaz,H., Salari,R., Sahinalp,S.C. and Backofen,R. (2009) A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, **25**, i365–i373.
10) Dimitrov,R.A. and Zuker,M. (2004) Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, **87**, 215–226.
11) Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
12) Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.
13) Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhöffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
14) Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
15) Huang,F.W.D., Qin,J., Reidys,C.M. and Stadler,P.F. (2009) Partition function and base pairing probabilities for RNA-RNA interaction prediction. *Bioinformatics*, **25**, 2646–2654.

Vol.2010-BIO-21 No.32
2010/6/19

16) Huang,F.W.D., Qin,J., Reidys,C.M. and Stadler,P.F. (2010) Target prediction and a statistical sampling algorithm for RNA-RNA interaction. *Bioinformatics*, **26**, 175–181.
17) Kato,Y., Akutsu,T. and Seki,H. (2009) A grammatical approach to RNA-RNA interaction prediction. *Pattern Recognit.*, **42**, 531–538.
18) Mückstein,U., Tafer,H., Hackermüller,J., Bernhart,S.H., Stadler,P.F. and Hofacker,I.L. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.
19) Pervouchine,D.D. (2004) IRIS: intermolecular RNA interaction search. *Genome Inform.*, **15**, 92–101.
20) Poolsap,U., Kato,Y. and Akutsu,T. (2009) Prediction of RNA secondary structure with pseudoknots using integer programming. *BMC Bioinform.*, **10**(Suppl 1), S38.
21) Poolsap,U., Kato,Y. and Akutsu,T. (2010) Dynamic programming algorithms for RNA structure prediction with binding sites. *Proc. Pac. Symp. Biocomput.*, **15**, 98–107.
22) Rehmsmeier,M., Steffen,P., Höchsmann,M. and Giegerich,R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.
23) Salari,R., Backofen,R.and Sahinalp,S.C. (2010) Fast prediction of RNA-RNA interaction. *Algorithms Mol.Biol.*, **5**, 5.
24) Vogel,J. and Wagner,E.G.H. (2007) Target identification of small noncoding RNAs in bacteria. *Curr. Opin. Microbiol.*, **10**, 262–270.