RactIP: fast and accurate prediction of RNA–RNA interaction using integer programming

Yuki Kato¹,*,†, Kengo Sato²,*,†, Michiaki Hamada³,⁴, Yoshihide Watanabe⁵, Kiyoshi Asai²,⁴ and Tatsuya Akutsu¹

¹Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, ²Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, ³Tosho Information & Research Institute, Inc, 2-3 Kanda-Nishichōki, Chiyoda-ku, Tokyo 101-8443, ⁴Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064 and ⁵Department of Mathematical Sciences, Faculty of Science and Engineering, Doshisha University, 1-3 Tataramiyakodani, Kyotanabe, Kyoto 610-0321, Japan

ABSTRACT

Motivation: Considerable attention has been focused on predicting RNA–RNA interaction since it is a key to identifying possible targets of non-coding 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 general, there is a trade-off between range of interaction type and efficiency of a prediction algorithm, and thus efficient computational methods for predicting comprehensive type of interaction are still awaited.

Results: We present RactIP, a fast and accurate prediction method for RNA–RNA interaction of general type using integer programming. RactIP can integrate approximate information on an ensemble of equilibrium joint structures into the objective function of integer programming using posterior internal and external base-pairing probabilities. Experimental results on real interaction data 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.

Availability: RactIP is implemented in C++, and the source code is available at http://www.ncrna.org/software/ractip/.

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Recent years have seen a renewal of interest in the biological roles of functional non-coding RNAs (ncRNAs). Modern studies have provided evidence that they act as ubiquitous regulators in living cells (Eddy, 2001; Vogel and Wagner, 2007). A class of small ncRNAs downregulates gene expression post-transcriptionally via base-pairing with target mRNAs to inhibit the translation into the corresponding proteins. Eukaryotic microRNAs (miRNAs) and small interfering RNAs (siRNAs) are very short and can integrate approximate information on an ensemble of equilibrium joint structures into the objective function of integer programming using posterior internal and external base-pairing probabilities. Experimental results on real interaction data 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.

To whom correspondence should be addressed.

© The Author(s) 2010. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
We demonstrate the prediction performance of IP. We make use of posterior probabilities such as respect to internal and external base pairs of two interacting RNAs. Integer Programming. In our IP formulation, the objective function, as well as using the Lagrangian relaxation. In our model, we employ additional constraints to represent stacking base pairs, which is expected to improve prediction accuracy.

The rest of the article is organized as follows. Section 2 describes our prediction model RactIP in detail after providing several preliminaries to grasp it. We show experimental results of interaction prediction and discuss them in Sections 3 and 4, respectively. Section 5 concludes the article.

2 METHODS

We propose a new method RactIP for RNA–RNA interaction prediction using 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.

It should be noted that the program RactIP actually solves the IP problem using libraries of a high-performance solver (see Section 3 for details).

2.1 Scoring functions for predicting RNA–RNA interaction

Let $\Sigma = \{A, C, G, U\}$ and $\Sigma'$ denote the set of all finite RNA sequences consisting of bases in $\Sigma$. For a sequence $a \in \Sigma^*$, let $|a|$ denote the number and binding site prediction, making a comparison with several state-of-the-art methods. Advantages of RactIP are summarized as follows:

- As for joint secondary structure prediction under the most general type of interaction, RactIP can run overwhelmingly faster than competitive prediction methods with $O(n^6)$ time complexities, where $n$ is the length of the longer input sequence. In fact, experimental results reveal that computation time of RactIP is an order of magnitude shorter than that of inteRNA (the joint structure prediction model) (Salari et al., 2010a), IntaRNA (the Loop Model) (Alkan et al., 2006) and rip (Huang et al., 2009, 2010). Recently, Salari et al. (2010b) proposed a sparsified dynamic programming algorithm whose time complexity is $O(n^5)$ on average. To the best of the authors’ knowledge, RactIP is the fastest method for predicting both internal structures and binding sites simultaneously on condition of the comprehensive class of interactions.

- RactIP is comparable in accuracy to inteRNA (the joint structure prediction model) and outperforms inteRNA and rip for joint structure prediction. From the viewpoint of binding site prediction that disregards (predicted) internal base pairs, accuracy of RactIP is as good as that of inteRNA (the binding site prediction model) and better than that of IntaRNA (Busch et al., 2008).

- The mathematical model of RactIP is compact. In particular, the IP objective function fit in well with the sum of the posterior probabilities that consider many complex loop energies necessary to achieve prediction of good quality.

- The IP-based method is flexible and extensible. Compared with other computational approaches, it is easy to change the model to cope with a desired class of secondary structures simply by adding or removing appropriate constraints. In our model, we employ additional constraints to represent stacking base pairs, which is expected to improve prediction accuracy.

Fig. 1. An example of RNA–RNA interaction containing two kissing hairpins. A broken line represents an internal base pair, and a black circle indicates a base that constitutes an external base pair (binding site).
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 \(a_{i+1}a_{i+2}\ldots a_{j}\in \Sigma^*\).

Let \(\mathcal{S}(a)\) be a space of all possible secondary structures of a sequence \(a\in \Sigma^*\). An element \(\sigma \in \mathcal{S}(a)\) is represented as a \(|a|\times |a|\) binary-valued triangular matrix \(x_{ij}\), where \(x_{ij}=1\) means that bases \(a_i\) and \(a_j\) form a base pair. We denote by \(P(x|a)\) the 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 \(\hat{\sigma} \in \mathcal{H}(a,b)\) is represented as a \(|a|\times |b|\) binary-valued matrix \(z_{ij}\), where \(z_{ij}=1\) means that base \(a_i\) interacts with the base \(b_j\). We denote by \(P(x|a,b)\) a probability distribution over \(\mathcal{H}(a,b)\). Let \(\mathcal{J}(S(a,b))\) 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, \(\mathcal{J}(S(a,b))\) is a subset of \(\mathcal{S}(a)\times \mathcal{S}(b)\times \mathcal{H}(a,b)\). We denote by \(P(x|a,b)\) a probability distribution over \(\mathcal{J}(S(a,b))\), where \(x=s_y\), \(y\in \mathcal{J}(S(a,b))\) such that \(x=s_y\), \(y\in \mathcal{H}(a,b)\) 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 structures (external pseudoknots) are disallowed.

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

**Problem 1** (RNA-RNA interaction prediction). Given two RNA sequences \(a=a[1..n] \in \Sigma^*\) (\(5'\rightarrow 3'\) direction) and \(b=b[1..m] \in \Sigma^*\) (\(3'\rightarrow 5'\) direction), predict a joint secondary structure \(\sigma \in \mathcal{J}(S(a,b))\) that maximizes the expected gain function \(G(\sigma,\theta)\) (see Supplementary Material for the derivation).

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

\[
G(\sigma,\theta) = G(s,\hat{s}) + G(\hat{s},\hat{\sigma}) + G(\hat{\sigma},\hat{\sigma}) \tag{1}
\]

where

\[
G(s,\hat{s}) = \sum_{i<j} [y_i \hat{y}_j = 1 | \hat{y}_j = 1 ] + [y_i \hat{y}_j = 0 | \hat{y}_j = 0 ] = 0, \tag{2}
\]

\[
G(\hat{s},\hat{\sigma}) = \sum_{i<j} [y_i \hat{y}_j = 1 | \hat{y}_j = 1 ] + [y_i \hat{y}_j = 0 | \hat{y}_j = 0 ] = 0, \tag{3}
\]

\[
G(\hat{\sigma},\hat{\sigma}) = \sum_{i<j} [y_i \hat{y}_j = 1 | \hat{y}_j = 1 ] + [y_i \hat{y}_j = 0 | \hat{y}_j = 0 ] = 0. \tag{4}
\]

Here, \(\text{condition}()\) is the indicator function that takes a value of 1 or 0 depending on whether the condition is true or false; \(y_i\) and \(\hat{y}_i\) are weight parameters for base pairs, and \(u\) 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:

\[
E_{G(a,b)}(\hat{\sigma}|a,b) = \sum_{\sigma \in \mathcal{J}(S(a,b))} G(\sigma,\theta) P_\theta(x|a,b). \tag{2}
\]

For the posterior distribution \(P_\theta(x|a,b)\) over a space of joint secondary structures, we can employ the p.l.RNA model (ChiuTse et al., 2009) and the x.t.p model (Huang et al., 2009, 2010). However, these exact models are impractical since \(O(n^3)\) 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:

\[
P_\theta(x|a,b) \propto P(x|a) P(y|b) P(z|a,b) \tag{3}
\]

Fig. 2. An illustration of the factorization (Equation (3)) of the posterior probability \(P(x|a,b)\) of a joint structure \(x\). A broken line shows an internal or external base pair.
is formed inside or outside the sequence (Fig. 4a). Internal pseudoknots and base can be paired with at most one base regardless of whether the base pair

half compared with internal base pairs. Constraints (6) and (7) mean that each pairs are less likely to be formed than internal ones. In this study, we fix

suggested by Alkan et al (2006), we set $\alpha \in (0, 1)$ so that external base pairs are less likely to be formed than internal ones. In this study, we fix $\alpha=0.5$, that is, external base pairs contribute to the objective function by half compared with internal base pairs. Constraints (6) and (7) mean that each base can be paired with at most one base regardless of whether the base pair is formed inside or outside the sequence (Fig. 4a). Internal pseudoknots and crossing interactions are prohibited by constraints (8), (9) and the constraint (10), respectively (Fig. 4b and c).

![Fig. 4. An illustration of several constraints of the IP formulation.](image)

Fig. 4. An illustration of variables used in the IP formulation. This variable setting enables the model to represent a kissing hairpin.

$$\text{maximize} \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij}(a) y_{ij} + \sum_{i=1}^{n} \sum_{k=1}^{m} p_{ik}(b) y_{ik} + \alpha \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij}$$

subject to

\[ x_{ij} + x_{ji} + x_{ik} + x_{ki} + x_{il} + x_{li} \leq 1 \quad (1 \leq i, j, k \leq n), \]

\[ x_{ij} + x_{ji} + x_{ik} + x_{ki} + x_{il} + x_{li} \leq 1 \quad (1 \leq i, j, k, l \leq m), \]

\[ x_{ij} + x_{ji} \leq 1 \quad (1 \leq i < j < k < l), \]

\[ y_{ij} + y_{ji} \leq 1 \quad (1 \leq i < j < k < l), \]

\[ z_{ij} + z_{ji} \leq 1 \quad (1 \leq i < j < k < l), \]

\[ x_{ij} + x_{ji} + x_{ik} + x_{ji} \leq 1 \quad (1 \leq i < j < k < l), \]

\[ x_{ij} + x_{ji} + x_{ik} + x_{ki} + x_{il} + x_{li} \leq 1 \quad (1 \leq i < j < k < l), \]

where $p_{ij}^{(a)}$, $p_{ik}^{(b)}$ 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.

Here, let us look more carefully into each equation of the above IP formulation. The objective function (5) shows an instantiation of the approximate estimator (4) using the base-pairing probabilities and hybridization probabilities. Note that the third term describing sum of scores of external base pairs is multiplied by a positive weight parameter $\alpha$. As suggested by Alkan et al. (2006), we set $\alpha \in (0, 1)$ so that external base pairs are less likely to be formed than internal ones. In this study, we fix $\alpha=0.5$, that is, external base pairs contribute to the objective function by half compared with internal base pairs. Constraints (6) and (7) mean that each base can be paired with at most one base regardless of whether the base pair is formed inside or outside the sequence (Fig. 4a). Internal pseudoknots and crossing interactions are prohibited by constraints (8), (9) and the constraint (10), respectively (Fig. 4b and c).

![Fig. 3. An illustration of variables used in the IP formulation. This variable setting enables the model to represent a kissing hairpin.](image)

Fig. 3. An illustration of variables used in the IP formulation. This variable setting enables the model to represent a kissing hairpin.

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 $\theta_i$ and $\theta_j$, respectively, defined in Section 2.1, before computing the optimal solution. As described in Section 2.3, this threshold cut is derived from the viewpoint of maximizing expected accuracy of joint structure prediction.

Incorporating stacked pairing constraints 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 for predicting secondary structure of a single RNA sequence proposed by Poolisap et al. (2009), we define further variables for promoting stacking base pairs as follows:

\[ x_{ij}^f = \begin{cases} 1 & \text{if base } a_i \text{ bonds with a base in position } i_j > i, \\ 0 & \text{otherwise}, \end{cases} \]

\[ x_{ij}^{f'} = \begin{cases} 1 & \text{if base } a_i \text{ bonds with a base in position less than } i, \\ 0 & \text{otherwise}. \end{cases} \]

In the IP formulation, we describe the definitions of $x_{ij}^f$ and $x_{ij}^{f'}$ as

\[ x_{ij}^f = \sum_{j=1}^{n} x_{ij} \quad (1 \leq i < n), \quad x_{ij}^{f'} = \sum_{j=1}^{n} x_{ij} \quad (1 < i \leq n). \]

We should notice that the following ingenious constraints containing linear combinations of these variables actually play a role in yielding stacking base pairs:

\[ x_{ij}^{f'} + (1-x_{ij}^{f'}) + x_{ij}^{f''} \geq 1 \quad (1 \leq i < n), \]

\[ x_{ij}^{f} + (1-x_{ij}^{f}) + x_{ij}^{f''} \geq 1 \quad (1 < i \leq n). \]

These constraints guarantee that if a base $a_i$ is paired with another one, the base(s) adjacent to $a_i$ must also form a base pair (Fig. 5). The rest of the variable definitions with respect to the sequence $b$ and interaction part, and related constraints, are similarly represented in the IP formulation.

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 IP problem. We employed the CONTRAfold model (Do et al., 2006).
Comparison with competitive methods for joint structure prediction

We first conducted experiments in joint secondary structure predictions of RNA binding sites. The first set comprises five pairs of RNA sequences with their joint structures predicted by RactIP\(^*\). The five RNA–RNA interaction pairs were predicted by RactIP\(^*\). The source code of RactIP\(^*\) is freely available at http://www.ncrna.org/software/ractip/.

### Table 1. Comparison with competitive methods for joint structure prediction

| Antisense-target | No. of sites | Sensitivity | PPV | F-measure | Time (s) |
|------------------|-------------|-------------|-----|-----------|----------|
|                  | RactIP      | inRNAs      | inteRNA | RactIP      | inRNAs      | inteRNA | RactIP      | inRNAs      | inteRNA | RactIP      | inRNAs      | inteRNA |
| CopA-CopT        | 2           | 1.000       | 1.000 | 0.731      | 0.754       | 0.846    | 0.655 | 0.860       | 0.917       | 0.691   | 0.13        |
| DIS-DIS          | 1           | 1.000       | 1.000 | 1.000      | 1.000       | 1.000    | 1.000 | 1.000       | 1.000       | 1.000   | 0.05        |
| IncRNAAs-RepZ    | 1           | 0.813       | 0.875 | 0.958      | 0.736       | 0.792    | 0.836 | 0.772       | 0.831       | 0.893   | 0.10        |
| Rinv-R2inv       | 1           | 1.000       | 0.900 | 0.800      | 1.000       | 0.900    | 0.889 | 1.000       | 0.900       | 0.842   | 0.03        |
| Tai-Tar\(^*\)    | 1           | 1.000       | 1.000 | 1.000      | 0.875       | 0.875    | 0.875 | 0.933       | 0.933       | 0.933   | 0.03        |
| Average          | 1           | 0.963       | 0.955 | 0.898      | 0.873       | 0.883    | 0.851 | 0.913       | 0.916       | 0.872   |             |

The five RNA–RNA interaction pairs were predicted by RactIP\(^*\), inRNAs (the joint structure prediction model) (Salari et al., 2010a) and inteRNA (the Loop Model) (Alkan et al., 2007). In the table, No. of sites represents the number of binding sites. We set the parameters for RactIP\(^*\) as \(a=0.5\), \(b=0.5\) and \(c=0.1\). 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 s for long sequences on Sun Fire X4600 2.6 GHz with 64 GB memory (Salari et al., 2010a).

We compared our method RactIP with two state-of-the-art methods: inRNAs (the exact model for joint structure prediction) (Salari et al., 2010a) and inteRNA (Alkan et al., 2006). 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 (Aksay \textit{et al}, 2007) (http://compbio.cs.sfu.ca/taverna/interna/). 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. It should be noted that computation time of RactIP includes both the pre-processing step to calculate posterior pairing probabilities and the one to solve the IP problem. We did not compare the running time strictly between the three methods due to difficulty in their availability. However, we would like to remark that Salari et al. (2010a) reported in their literature that inRNAs runs for ~4000 s on Sun Fire X4600 2.6 GHz with 64 GB memory to predict the joint structures of CopA-CopT and IncRNAAs-RepZ, respectively. Meanwhile, RactIP consumes only 0.13 s and 0.10 s to predict the joint structures of CopA-CopT and IncRNAAs-RepZ, respectively, on Mac OS X 10.6 running on Intel Core 2 Duo 2.13 GHz with 2 GB memory.

### 3.2 Data

In our experiments, we used two datasets of RNA–RNA interactions. The first set comprises five pairs of RNA sequences with their joint secondary structures including kissing hairpins, which was used by Kato \textit{et al} (2009). We made use of this set to evaluate the performance of joint secondary structure prediction. The second set contains 18 sRNA–target pairs with their binding sites, which was used by Busch \textit{et al} (2008). All of the two datasets (i.e. 23 pairs of interacting RNAs) were used to assess the performance of binding site prediction.

### 3.3 Joint structure prediction

We first conducted experiments in joint secondary structure prediction on the dataset compiled by Kato \textit{et al} (2009). 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:

\[
F = \frac{2 \times \text{sensitivity} \times \text{PPV}}{\text{sensitivity} + \text{PPV}}
\]

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 \(p_{ij}\) and \(q_{ij}\) for RNAduplex, which is a program from the Vienna RNA package (Hofacker \textit{et al}, 1994; Hofacker, 2003) 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, designing a forward-backward-like algorithm. The source code of RactIP is freely available at http://www.ncrna.org/software/ractip/.

### 3.4 Binding site prediction

In the second experiment, we assessed the performance of predicting binding sites on the dataset compiled by Kato \textit{et al} (2009) and Busch \textit{et al} (2008). 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) (Salari \textit{et al}, 2010a) extracted from their literature, and inteRNA (Busch \textit{et al}, 2008) with default settings, 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 inteRNA and inRNAs can consider only one or two accessible regions that are putative binding sites.

### 3.5 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 accuracy for the joint secondary structure prediction model (Salari \textit{et al}, 2010a) and inteRNA (Alkan \textit{et al}, 2006).
The four RNA–RNA interaction pairs were predicted by \textit{RactIP} with base-pairing probabilities calculated by \textit{rip} \cite{Huang2009,Huang2010a}. We set the parameters for \textit{RactIP} as $\alpha = 0.5$, $\eta_0 = 0.3$ and $\eta_1 = 0.2$. Running time of \textit{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 \textit{rip} was measured on Linux kernel 2.6.30 running on Intel Xeon 3.33 GHz with 32 GB memory.

### Table 3. Comparison of accuracy and running time for joint structure prediction

| Antisense target | Sensitivity | F-measure | Time (s) |
|------------------|-------------|-----------|----------|
|                  | \textit{RactIP} \textit{rip} \textit{rip+RactIP} | \textit{RactIP} \textit{rip} \textit{rip+RactIP} | \textit{RactIP} \textit{rip} \textit{rip+RactIP} |
| DIS-DIS          | 1.000       | 1.000     | 1.000     |
| \textit{InRNA} \textit{A-RegZ} | 0.131       | 0.131     | 0.131     |
| \textit{Rinv-R2inv} | 1.000       | 1.000     | 1.000     |
| Tar-Tar*         | 1.000       | 1.000     | 1.000     |

The prediction accuracy of the factorized model that we proposed with those of the naive model by \textit{rip} \cite{Huang2009,Huang2010a}. \textit{rip} calculates exact base-pairing probabilities of internal base pairs and external base pairs by taking $O(n^2)$ space where $n$ is the length of the longer sequence. We compared \textit{RactIP} with \textit{rip}, which samples joint structures from internal and external base-pairing probabilities, and \textit{RactIP} combined with \textit{rip}, in which internal and external base-pairing probabilities calculated by \textit{rip} were used in the \textit{IP} \cite{Huang2009,Huang2010a} instead of factorized ones. Note that \textit{rip} failed to calculate base-pairing probabilities for the CopA-CopT pair since their length might be too long for \textit{rip}. As shown in Table 3, our approximation by factorization is significantly faster than the naive 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 DISCUSSION

We employed the threshold cut technique to reduce the search space for the optimal joint secondary structure, which makes \textit{RactIP} run much faster than existing state-of-the-art algorithms for joint structure prediction. Let us stress again that there is a close relation between the threshold cut and maximizing expected accuracy.
We proposed RactIP and would like to emphasize again that makes our method fail to optimize appropriately. Nevertheless, inRNAs work, optimizing hybridization scores appropriately is necessary to framework of the integer programming formulation. As our future complex interactions with more than one binding site in the pairing probabilities and hybridization probabilities. The main (e.g. the accuracy decreased when adopting the identical scoring scheme a joint structure shown in Equation (3). In fact, prediction scheme due to the factorization of the posterior probability of objective function. It is possible to adopt such a hybrid scoring for more details.

The results shown in Table 2 tell us that RactIP performs worse than IntaRNA and is not significantly better than IntaRNA on CopA–CoP and OxyS–thlA pairs with two binding sites. One reason is that lack of interaction data with multiple binding sites makes our method fail to optimize appropriately. Nevertheless, we would like to emphasize again that RactIP can deal with complex interactions with more than one binding site in the framework of the integer programming formulation. As our future work, optimizing hybridization scores appropriately is necessary to improve prediction performance on the data with multiple binding sites.

5 CONCLUSION

We proposed RactIP, a novel method for predicting RNA–RNA interaction of general type using IP. In our approach, the threshold cut technique was adopted to reduce the complexity of the solution space of the IP problem, which also leads to maximizing expected accuracy. Experimental results on real interaction data 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. Although it is difficult to evaluate theoretically the time complexity of our IP-based approach, experimental validations revealed that RactIP can run much faster than competitive methods for predicting joint secondary structures. This is an important fact to stress since RactIP is expected to improve prediction performance in unknown target search in long genomes by predicting respective intramolecular structures as well as intermolecular binding sites in practical time. For this purpose, we should also show that RactIP can discriminate between targets and non-targets, which is left as our future work. RactIP not only achieved success in RNA–RNA interaction prediction but also showed further possibility of applying the fast IP-based method with threshold cut to other biologically important problems, which are worthwhile and challenging tasks.

ACKNOWLEDGEMENTS

The authors would like to thank Ms Uynane Pool sap 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.

Funding: Young Scientists (B) (KAKENHI) from Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan (#22700313 to Y.K., #22700305 to K.S.).

Conflict of Interest: none declared.

REFERENCES

Aksay,C. et al. (2007) taveRNA: a web suite for RNA algorithms and applications. Nucleic Acids Res., 35, W325–W329.

Alix,A.C. et al. (2006) RNA–RNA interaction prediction and antisense RNA target search. J. Comput. Biol., 13, 267–282.

Andronescu,M. et al. (2005) Secondary structure prediction of interacting RNA molecules. J. Mol. Biol., 348, 987–1001.

Bauer,M. et al. (2007) Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. BMC Bioinformatics, 8, 271.

Bernhart,S.H. et al. (2006) Partition function and base pairing probabilities of RNA heterodimers. Algorithm Mol. Biol., 1, 3.

Brent,S. (2002) Antisense-RNA regulation and RNA interference. Biochim. Biophys. Acta, 1575, 15–25.

Brudno,M. et al. (2002) RNA loop-loop interactions as dynamic functional motifs. Bioinformatics, 18, 925–944.

Busch,A. et al. (2008) IntaRNA: efficient prediction of bacterial RNA targets incorporating target site accessibility and seed regions. Bioinformatics, 24, 2449–2456.

Chitaia,H. et al. (2009) A partition function algorithm for interacting nucleic acid strands. Bioinformatics, 25, 1365–1373.

Dimitrov,R.A. and Zucker,M (2004) Prediction of hybridization and melting for double-stranded nucleic acids. Biophys. J., 87, 215–226.

Do,C.B. et al. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics, 22, e90–e98.

Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet., 2, 919–929.

Hamada,M. et al. (2009) Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics, 25, 465–473.

Hofacker,I.L. et al. (1994) Fast folding and comparison of RNA secondary structures. Monatsh. Chem., 125, 167–188.

Hofacker,I.L. et al. (2003) Vienna RNA secondary structure server. Nucleic Acids Res., 31, 3429–3431.

Huang,F.W.D. et al. (2009) Partition function and base pairing probabilities for RNA–RNA interaction prediction. Bioinformatics, 25, 2686–2694.

Huang,F.W.D. et al. (2010) Target prediction and a statistical sampling algorithm for RNA–RNA interaction. Bioinformatics, 26, 175–181.

Kato,Y. et al. (2009) A grammatical approach to RNA–RNA interaction prediction. Pattern Recognit., 42, 531–538.

Mickelstein,U. et al. (2006) Thermodynamics of RNA–RNA binding. Bioinformatics, 22, 1177–1182.

Pervouchine,D.D. (2004) IRIS: intermolecular RNA interaction search. Nucleic Acids Res., 32, 262–270.

Peters,S.H. et al. (2009) A grammatical approach to RNA-RNA interaction prediction. Genome Inform., 15, 92–101.

Poolap,U. et al. (2009) Prediction of RNA secondary structure with pseudoknots using integer programming. BMC Bioinformatics, 10 (Suppl. 1), S38.

Poolap,U. et al. (2010) Dynamic programming algorithms for RNA-structure prediction with binding sites. Proc. Pac. Symp. Biocomput., 15, 98–107.

Rémésy,M. et al. (2004) Fast and effective prediction of microRNA/target duplexes. RNA, 10, 1507–1517.

Salan,R. et al. (2010) Fast prediction of RNA–RNA interaction. Algor. Mol. Biol., 5, 5.

Salan,R. et al. (2010b) Time and space efficient RNA-RNA interaction prediction via sparse folding. Lect. Notes Bioinformatics, 6044, 473–490.

Vogel,J. and Wagner,E.G.H. (2007) Target identification of small noncoding RNAs in bacteria. Curr. Open. Microbiol., 10, 262–270.

Wagner,E.G.H. and Fährd,R.K. (2002) Antisense RNAs everywhere? Trends Genet., 18, 223–226.