Predictions of RNA secondary structure by combining homologous sequence information

Michiaki Hamada1,2,3,∗, Kengo Sato2,4, Hisanori Kiryu2, Toutai Mituyama2 and Kiyoshi Asai2,5
1Mizuho Information & Research Institute, Inc, 2–3 Kanda-Nishikicho, Chiyoda-ku, Tokyo 101–8443, 2Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2–41–6, Aomi, Koto-ku, Tokyo 135–0064, 3Department of Computational Intelligence and System Science, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama 226–8503, 4Japan Biological Informatics Consortium (JBIIC), 2–45 Aomi, Koto-ku, Tokyo 135–8073 and 5Graduate School of Frontier Sciences, University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa 277–8562, Japan

ABSTRACT
Motivation: Secondary structure prediction of RNA sequences is an important problem. There have been progresses in this area, but the accuracy of prediction from an RNA sequence is still limited. In many cases, however, homologous RNA sequences are available with the target RNA sequence whose secondary structure is to be predicted.

Results: In this article, we propose a new method for secondary structure predictions of individual RNA sequences by taking the information of their homologous sequences into account without assuming the common secondary structure of the entire sequences. The proposed method is based on posterior decoding techniques, which consider all the suboptimal secondary structures of the target and homologous sequences and all the suboptimal alignments between the target sequence and each of the homologous sequences. In our computational experiments, the proposed method provides better predictions than those performed only on the basis of the formation of individual RNA sequences and those performed by using methods for predicting the common secondary structure of the homologous sequences. Remarkably, we found that the common secondary predictions sometimes give worse predictions for the secondary-structure of a target sequence than the predictions from the individual target sequence, while the proposed method always gives good predictions for the secondary structure of target sequences in all tested cases.

Availability: Supporting information and software are available online at: http://www.ncrna.org/software/centroidfold/ismb2009/.

Contact: hamada-michiaki@aist.go.jp

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
Secondary structure prediction of RNA sequences is one of the most important fundamental problems in biological sequence information analysis. The importance of accurate predictions has increased due to the recent findings in functional non-coding RNAs. Although a number of algorithms and tools have been developed for this problem (e.g. Andronescu et al., 2007; Do et al., 2006a; Hamada et al., 2008; Hofacker et al., 1994; Parisien and Major, 2008), their accuracies are limited. In real cases, however, it is expected that the accuracies can be improved by taking into account the information of RNA sequences that are homologous to the target RNA sequence whose secondary structure is to be predicted. There have been proposed several methods with this respect (e.g. Hamada et al., 2006; Hofacker et al., 2002; Kiryu et al., 2007). The problem targeted by this article is defined as follows.

Problem 1. Given a single RNA sequence x and a set of its homologous sequences D, predict a secondary structure of x using the information of D.

It is assumed that x and each sequence x′ in D share a common secondary structure, and the secondary structure of x is predicted by combining information present in D. This problem can be solved by mapping the predicted common structure of the entire sequences to the target sequence as follows: (i) compute a multiple alignment A for the set of sequences D∪{x} (e.g. Do et al., 2005, 2008; Tabei et al., 2008); (ii) predict a common secondary structure of A (e.g. Bernhart et al., 2008; Hamada et al., 2008; Seemann et al., 2008) and (iii) predict a secondary structure of x by mapping the predicted common secondary structure to the target RNA sequence x. The obtained secondary structures are more reliable in general than the predictions only from individual sequences. There are problems in the above approach, however, because the secondary structures of individual sequences are not exactly the same as the common secondary structure and the quality of the multiple alignment strongly influences the performance of the prediction.

In this article, we propose a novel estimator for Problem 1 which yields a direct prediction of the secondary structure of x by combining the information of the homologous sequences D without assuming the multiple alignment or the common structure of the entire sequences (Fig. 1). The proposed method models the probabilistic distribution of the pairwise common structure between the target RNA sequence and each homologous sequence, and maximizes the expected accuracy of the prediction of the target RNA sequence under the combined probabilistic distribution. The expectation is calculated under all the suboptimal secondary structures of the target sequence and under all the suboptimal pairwise alignments of the target sequence and its homologous sequences using posterior decoding method, which is based on a posteriori probabilities such as base-pairing probabilities of RNA or match probabilities of alignment and has been successfully applied in bioinformatics (Bradley et al., 2008; Carvalho and Lawrence, 2008).
No use of $D$

\[ \begin{align*}
\text{et al.} = \{ \theta \} \sum_{\theta} \text{following constraints: (i) } p \text{ obtain a probability distribution} \\
\text{S} \end{align*} \]

\[ \begin{align*}
\text{METHODS} \\
\text{2 METHODS} \\
\text{2.1 Discrete spaces and probability distributions} \\
\text{In this article, the following discrete spaces and probability distributions on those spaces play important roles in order to define proposed estimators.} \\
\text{2.1.1 A space of secondary structures: } \text{S(x)} \\
\text{A secondary structure of an RNA sequence } x \text{ is represented as an upper triangular binary matrix } \\
\theta = \{ \theta_{ij} \}, \text{ where the } (i,j) \text{ element } \theta_{ij} = 1 \text{ if } x_i \text{ and } x_j \text{ form a base pair and to } 0 \text{ if } x_i \text{ and } x_j \text{ do not form a base pair (Fig. 2).} \\
\text{A consistent secondary structure represented by } \theta \text{ should satisfy the following constraints: (i) } \sum_{x_i} \theta_{ij} + \sum_{x_j} \theta_{ij} \leq 1 \text{ for any } i, \text{ (ii) } \theta_{ij} + \theta_{ji} \leq 1 \text{ for any base pairs which results in a pseudo-knot is not allowed).} \\
\text{We denote with } \text{S(c)} \text{ the space of all secondary structures } \theta \text{ of an RNA sequence } x. \\
\text{Note that to predict the secondary structure } \theta \text{ is equivalent to predicting a point } \theta \text{ in } \text{S(x).} \\
\text{We obtain a probability distribution } p^{\text{PHMM}}(x) \text{ on } \text{S(x)} \text{ by using the McCaskill model (McCaskill, 1990), the CONTRAfold model (Do et al., 2006a), the Simfold model (Andreeva et al., 2007) and the SCFG model (Dowell and Eddy, 2004).} \\
\text{By using the distributions, a base-pairing probability } p_{ij}^{\theta(x,x')}, \text{ which is the probability with which } x_i \text{ and } x_j \text{ form a base pair, is computed as} \\
p_{ij}^{\theta(x,x')} = p^{\text{PHMM}}(x) = \frac{1}{|x|} \sum_{x_i} \sum_{x_j} \theta_{ij} = 1 \text{ if a base pair } (x_i, x_j) \text{ is included in } \theta. \\
\text{A structural alignment (e.g. Kiryu et al., 2007a) A structural alignment } (\theta, x') \text{ is denoted as} \\
\text{SA}(\theta, x') \\
\text{A space of pairwise alignments: } \text{A}(x, x') \\
\text{Similarly to the secondary structure of an RNA sequence, a pairwise alignment of two sequences } x \text{ and } x' \text{ (e.g. DNA sequences, RNA sequences or amino acid sequences) is represented as a binary matrix } \theta = \{ \theta_{ij} \}, \text{ where } \theta_{ij} = 1 \text{ if } x_i \text{ is aligned with } x'_j \text{ and } \theta_{ij} = 0 \text{ if } x_i \text{ is not aligned with } x'_j (Fig. 3). \\
\text{A consistent alignment represented by } \theta \text{ should satisfy the following constraints: (i) } \sum_{x_i} \theta_{ij} \leq 1 \text{ for any } i, \text{ (ii) } \sum_{x_j} \theta_{ij} \leq 1 \text{ for any } j \text{ and } i < j \text{ in this article, the space of all pairwise alignments of } x \text{ and } x' \text{ is denoted by } \text{A}(x, x'). \\
\text{We can obtain the probability distribution } p^{\text{PHMM}}(\theta(x,x'), x') \text{ by using the PHMM model (Durbin et al., 1998), the Miyazawa model (Miyazawa, 1995), the ProbAlign model (Roshan and Livesay, 2006), the CONTRAAlign model (Do et al., 2006b).} \\
\text{By using the distributions, the alignment match probability } p^{\theta(x,x')}_{ij}, \text{ which is the probability that } x_i \text{ is aligned with } x'_{j}, \text{ is computed as } \\
p^{\theta(x,x')}_{ij} = p^{\text{PHMM}}(\theta_{ij} = 1) = \sum_{x_i} \sum_{x_j} \theta_{ij} = 1 \theta_{ij} = 1 \theta_{ij} \text{ in a loop region of } x. \\
\text{Note that } p^{\theta(x,x')}_{ij} \mid_{i \neq j} \text{ can be computed by using a variant of the forward-backward algorithm whose time complexity is equal to } O(|x|^2 t^2). \\
\text{2.1.3 A space of structural pairwise alignments of RNA sequences: } \text{S}(\text{A}(x, x')) \\
\text{A structural alignment (e.g. Kiryu et al., 2007a; Sankoff, 1995) of two RNA sequences } x \text{ and } x' \text{ is represented as } \theta = \{ \theta_{ij} \}, \text{ where } \theta_{ij} = 1 \text{ if a base pair } (x_i, x_j) \text{ is included in } \text{A}(x, x') \text{ (Fig. 4).} \\
\text{In accordance with a consistent structural alignment, each element in } \theta \text{ cannot take an arbitrary value in } \{0, 1\}, \text{ where } \theta \text{ satisfies several conditions.} \\
\text{Moreover, a space of structural alignments of two RNA sequences } x \text{ and } x' \text{ is denoted as } \text{S}(\text{A}(x, x')). \\
\text{We obtain a probability distribution } p^{\text{PHMM}}(\theta(x,x'), x') \text{ by the pair SCFG model, the Sankoff model (Sankoff, 1985).} \\
\text{We also denote a probability that a base pair } (x_i, x_j) \text{ in } x \text{ is aligned with a base pair } (x'_i, x'_j) \text{ in } x' \text{ as } p^{\theta(x,x')}_{ij} = p^{\text{PHMM}}(\theta_{ij} = 1 | x, x'). \\
\text{Note that } p^{\theta(x,x')}_{ij} \mid_{i \neq j} \text{ can be computed by using a variant of the forward-backward algorithm whose time complexity is equal to } O(|x|^2 t^2). \\
\text{Fig. 2. Representation of a secondary structure of an RNA sequence.} \\
\text{Fig. 3. Representation of a pairwise alignment of two RNA sequences.}
computed by using a variant of the inside–outside algorithm of the Sankoff model (Sankoff, 1985), whose time complexity is $O(|x|^3|y|^3)$.

### 2.2 Designing estimators for Problem 1

Most of the existing decoding methods are regarded as the following estimators (e.g. Carvalho and Lawrence, 2008; Hamada et al., 2008): if $I$ is a space from which we would like to obtain a prediction, referred to as a predictive space, $\Theta$ is a space which is referred to as a parameter space and is potentially different from the predictive space, $p(\theta|d)$ is a probability distribution on the parameter space $\Theta$ given a dataset $d$, and $G(\theta, y)$ is a gain function between $\Theta$ and $I$, then an estimator on the predictive space is defined as

$$\hat{y} = \arg \max_{y \in I} E_{\theta}(G(\theta, y))$$

(cf. Ding et al., 2005; Do et al., 2006; Hamada et al., 2008; Kiryu et al., 2007b). Here, we can design a superior estimator by defining the gain function and the parameter space appropriately. For example, (Hamada et al., 2008) have proposed an estimator which maximizes the expectation $\alpha_T h_T(a_T)$ over a space $A_T$, with respect to a probability distribution on $S(x)$, and confirmed that their estimators are superior to the MEA estimator used in CONTRAfold (Do et al., 2006a). Regarding Problem 1, we would like to predict the secondary structure of $x$ so that the predictive space $I$ is $S(x)$. Therefore, it is necessary to define the parameter space $\Theta$ and the probability distribution $p(\theta|x)$ on the parameter space $\Theta$.

#### 2.2.1 Estimator 1

Our assumption in Problem 1 is that $x$ and every $y \in D$ share a common secondary structure. The common secondary structure is naturally represented as a structural alignment of $x$ and $y$. (See Section 2.1.3). Therefore, we set the parameter space as $\Theta^0 = \prod_{i \in \text{alignment}} \Theta^0(s_i)$, and the probability distribution on $\Theta^0$ is $p(\theta^0|x, D) = \prod_{i \in \text{alignment}} p(\theta^0(s_i)|x, D)$, where $\theta^0$ is a structural alignment of $x$ and $y$. (i.e. $\Theta^0(s_i, x)$). Moreover, the gain function is defined by $G^{(0)}(\theta^0, y) = \sum_{i \in \text{alignment}} G^0_i(\theta^0(s_i), y)$, for a prediction $y \in S(x)$, where $|D|$ denotes the number of sequences in $D$ and $\alpha > 0$ is a parameter which adjusts the sensitivity and the positive predictive values (PPV) (see Section 3 for definition) of a predicted secondary structure. In this probability $\frac{1}{|D|} \sum_{i \in \text{alignment}} G^0_i(\theta^0(s_i), y)$ is equal to the averaged number of a pair $(s_i, s_j)$ which does not form a base pair in the set. Hence, the more base pairs $(s_i, s_j)$ there exist in the set of structural alignments, the more gain in the gain function is given for the prediction $y_0 = 1$. Then, we define the estimator which maximizes the expectation of the gain function $G^{(0)}(\theta^0, y)$ on the probability distribution $p(\theta^0|x, D)$ as

$$\hat{y} = \arg \max_{y \in S(x)} \sum_{i \in \text{alignment}} G^0_i(\theta^0(s_i), y)$$

It is clear that the estimator is equivalent to the $\gamma$-centred estimator (Hamada et al., 2008) on $S(x)$ when the probability distribution is defined by $p(\theta|x, D) = \frac{1}{|D|} \sum_{i \in \text{alignment}} p(\theta(s_i)|x, D)$, where $\theta$ is the natural map from a structural alignment $\theta' \in S(A(x, s))$ to the secondary structure $\theta \in S(x)$. The optimal secondary structure is computed by a Nussinov-type dynamic programming (Nussinov et al., 1978) as follows:

$$M_{i,j} = \max \begin{cases} M_{i+1,j}, & \text{if } x_i \neq x_j, \\ M_{i,j-1}, & \text{if } y_{ij} = 0 \\ M_{i,j-1} + \gamma (|x| + |y|), & \text{otherwise} \end{cases}$$

where $p_0 = \frac{1}{|D|} \sum_{i \in \text{alignment}} p(\theta(s_i)|x, D)$. As described in Section 2.1.3, the time complexity to obtain $\sum_{i \in \text{alignment}} p(\theta(s_i)|x, D)$ is $O(|x|^3|y|^3)$, so it requires $O(|D|)|x|^3|y|^3$ time for computing $p_0$. The time complexity is required for the dynamic programming (3) in the case of $|y| > 1$. Therefore, the total time complexity is $O(|D|)|x|^3|y|^3$. This computational cost is too large to implement a practical application, so we approximate the estimator in the following sections.

#### 2.2.2 Estimator 2

Since $S(A(x, s))$ and a probability distribution on the space require huge computational cost as noted in the previous section, we replace the parameter space $\Theta^0$ and the gain function $G^{(0)}(\theta^0, y)$ by their approximations with some heuristics. The parameter space is approximated by the factorization of $\Theta^0$ such that $\Theta^0 = S(x) \times \prod_{i \in \text{alignment}} \Theta^0(s_i, x)$ and the corresponding probability distribution on $\Theta^0$ is defined as

$$p(\theta^0|x, D) = \prod_{i \in \text{alignment}} p(\theta^0(s_i)|x, D) p(\theta^0(x))$$

for $\theta = (\theta^0(x), \prod_{i \in \text{alignment}} \theta^0(s_i, x))$, where $\theta^0(x) \in S(x)$, $\theta^0(s_i, x)$ and $\theta^0(s_i, x)$ exist in $S(A(x, s))$. We also consider two kinds of approximations of $\sum_{i \in \text{alignment}} p(\theta^0(s_i)|x, D)$ in the gain function $G^{(0)}(\theta^0, y)$ as follows:

$$\sum_{i \in \text{alignment}} p(\theta^0(s_i)|x, D) = \sum_{i \in \text{alignment}} p(\theta^0(s_i)|y_0 = 1) + (1 - \alpha) \sum_{i \in \text{alignment}} p(\theta^0(s_i)|y_0 = 1) = 1 \frac{1}{|D|} \sum_{i \in \text{alignment}} p(\theta^0(s_i)|y_0 = 1) + 1 \frac{1}{|D|} \sum_{i \in \text{alignment}} p(\theta^0(s_i)|y_0 = 1)$$

and

$$\sum_{i \in \text{alignment}} p(\theta^0(s_i)|x, D) = \sum_{i \in \text{alignment}} p(\theta^0(s_i)|y_0 = 1) + (1 - \alpha) \sum_{i \in \text{alignment}} p(\theta^0(s_i)|y_0 = 1) = 1 \frac{1}{|D|} \sum_{i \in \text{alignment}} p(\theta^0(s_i)|y_0 = 1) + 1 \frac{1}{|D|} \sum_{i \in \text{alignment}} p(\theta^0(s_i)|y_0 = 1)$$

and

$$G^{(0)}(\theta^0, y) = \sum_{i \in \text{alignment}} G^0_i(\theta^0(s_i), y)$$

$G^{(0)}_0(\theta^0(s_i), y) = p(\theta^0(s_i)|y_0 = 1) + \alpha p(\theta^0(s_i)|y_0 = 0)$

and

$$G^{(0)}(\theta^0, y) = \sum_{i \in \text{alignment}} G^0_i(\theta^0(s_i), y)$$

$G^{(0)}_0(\theta^0(s_i), y) = p(\theta^0(s_i)|y_0 = 1) + \alpha p(\theta^0(s_i)|y_0 = 0)$

respectively. Then, we introduce two estimators in order to maximize the expectation of $G^{(0)}(\theta^0, y)$ at $G^{(2, 2)}(\theta^0, y)$ under the probability distribution.
p^{2S}(\theta, |D|). We can obtain the secondary structure of each estimator by replacing $p^{0}_{x}$ with

$$p^{(2,1)}_{y} = p^{0}_{x} + \frac{1}{|D|} \sum_{y \in \mathcal{Y}} p^{(3,1)}_{x,y},$$

and

$$p^{(2,2)}_{y} = \frac{1}{|D|} \sum_{y \in \mathcal{Y}} p^{(3,2)}_{x,y},$$

in Equation (3), respectively. We refer to these two estimators as ‘appro2-1’ and ‘appro2-2’, respectively. As described in Section 2.1.2, the estimator $p^{(2,2)}_{x}$ is used in an attempt to reduce the computational cost. This estimator is still too large for calculating secondary structures whose length is more than several hundred bases. Therefore, we approximate this estimator in an attempt to reduce the computational cost.

2.2.3 Estimator 3 (proposed estimator) By approximating $p^{(3,1)}_{x,y}$ to $p^{0}_{x,y}$, we replace $p^{(2,1)}_{y}$ and $p^{(2,2)}_{y}$ in the previous section by

$$p^{(3,1)}_{y} = p^{0}_{y} + \frac{1}{|D|} \sum_{x \in \mathcal{X}} p^{(3,1)}_{x,y},$$

and

$$p^{(3,2)}_{y} = \frac{1}{|D|} \sum_{x \in \mathcal{X}} p^{(3,2)}_{x,y},$$

respectively. Note that $p^{(3,1)}_{x,y}$ and $p^{(3,2)}_{x,y}$ take a value between 0 and 1, although they are not probabilities (more precisely, $p^{(3,1)}_{x,y}$ cannot be obtained by any probability distribution on $\mathcal{Y}(x)$). We refer to these two estimators as ‘appro3-1’ and ‘appro3-2’, respectively. As $\alpha = 1/|D| + 1$, $p^{(3,1)}_{x,y}$ is the same as a probabilistic consistency transformation of the base pairing probability matrix in Kiryu et al. (2007a). Even if we employ these approximations, the time complexity to obtain $p^{(3,1)}_{y}$ or $p^{(3,2)}_{y}$ remains $O(|D|^3 max_{x \in \mathcal{X}} \mathcal{Y}(x)^{2})$. However, since the number of base pairs $(i', j')$ in $x'$ where $p^{(3,1)}_{i', j'} > \delta$ for a threshold $\delta$ is regarded as $O(|c|)$ for a constant $c$, $p^{(3,1)}_{y}$ and $p^{(3,2)}_{y}$ can be computed by $O(|D|^2 max_{x \in \mathcal{X}} \mathcal{Y}(x)^{2})$ if we sum over the base pairs $(i', j')$ where $p^{(3,1)}_{i', j'} > \delta$ in the equation $G_{\theta}(x)$ and $p^{(2,2)}_{y}$. By definition, we see that the proposed estimators consider all suboptimal secondary structures of $x$ and $x'$ as the cases when $\theta = 1$.

**Fig. 5.** The difference between two gain functions $G_{\theta}^{(2,1)}(x, y)$ and $G_{\theta}^{(2,2)}(x, y)$. Each row indicates a configuration of $\theta$ on the target sequence $x$ and each homologous sequence $x'$, and its positive contributions to the gain functions for $\gamma = 1$. For example, the middle row means that the contribution of $\theta_{i}^{(3,1)} = 1$ and $\theta_{j}^{(3,1)} = 0$ has a positive (1-\alpha) contribution for $\eta_{i,j} = 1$ in the gain function $G_{\theta}^{(2,1)}$, while have no contribution in the gain function $G_{\theta}^{(2,2)}$.

3 EXPERIMENTS

We conducted all experiments in this section on our Linux cluster machines, each of which has 2 GHz CPU of AMD Opteron(tm) Processor 246 and 4 GB memory. The accuracy of predicted secondary structures was evaluated through the following standard evaluation measures: Sensitivity = TP/(TP+FN) and PPV = TP/(TP+FP), where TP is the number of correctly predicted base pairs, FN is the number of base pairs in the reference structure that were not predicted and FP is the number of incorrectly predicted base-pairs. In all the figures described below, we plot the curve at $\gamma \in \{2^k : -5 \leq k \leq 10, k \in \mathbb{Z}\} \cup \{6\}$ for the proposed estimators, the averaged $\gamma$-centroid estimators and the $\gamma$-centroid estimators (Hamada et al., 2008), which are implemented in the CentroidFold software.

3.1 Dataset

We used three datasets in our experiments. A summary of our datasets is described in Table 1.

3.1.1 The dataset1 We created a dataset called ‘dataset1’ by utilizing the dataset in Kiryu et al. (2007b) which contains 85 reference alignments of 10 sequences taken from 17 RNA families in the Rfam database (Griffiths-Jones et al., 2005) as follows: we selected each RNA sequence in each alignment in the Kiryu’s dataset as a target sequence $x$ and the other sequences in the alignment as homologous sequences $D$. The reference structure of the target sequence $x$ is given by mapping the consensus structure in the alignment to the target sequence. As a result, we obtained 850 test data, each of which contains a target sequence and nine homologous sequences. Note that the dataset in Kiryu et al. (2007b) contained only the manually curated seed alignments with the consensus structures published in literature.

3.1.2 The dataset2 and dataset3 We created dataset called ‘dataset2’ and ‘dataset3’ from the RNA STRAND database.

**Table 1.** Summary of dataset used in our experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Data</th>
<th>#Homologous (D)</th>
<th>Length of sequence (bp)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>dataset1</strong></td>
<td>850</td>
<td>9</td>
<td>48–389</td>
<td>Kiryu et al. (2007b)</td>
</tr>
<tr>
<td><strong>dataset2</strong></td>
<td></td>
<td></td>
<td>50–500</td>
<td>Andronescu et al. (2008)</td>
</tr>
<tr>
<td><strong>dataset3</strong></td>
<td>215</td>
<td>2–20</td>
<td>500–1500</td>
<td>Andronescu et al. (2008)</td>
</tr>
</tbody>
</table>

#data means the number of test data which contains a target sequence and several homologous sequences. #homologous means the number of homologous sequence included in each test data. See Section 3.1 for details of each dataset.
secondary structure to the target sequence. As a result, the numbers of test data in the dataset2 and the dataset3 are 1547 and 215, respectively.

### 3.2 Compared methods

In the following sections, we compared the proposed method with three types of methods (Fig. 1) for Problem 1 as follows. Note that the default parameters were used in each tool unless the parameters were specified. (M1) Predict a secondary structure of a target RNA sequence \( x \) without using homologous sequence information. In this method, we used the tools of secondary structure prediction from an individual RNA sequence: RNAfold version 1.8.1 (Hofacker et al., 1994), SimFold in MultiRNAfold-1.11 with the ISMB 2007 best parameters (Andronescu et al., 2007) and CentroidFold version 0.0.4 (Hamada et al., 2008) with the CONTRAfold model and do not require fixed alignments. (M2) Align \( [x] \cup D \) and predict a common secondary structure of the predicted multiple alignment, and then mapped the common secondary structure to the target sequence \( x \). In this method, we used the tools of the multiple alignment of RNA sequences: ProbCons (RNA) (Do et al., 2005), RAf v1.00 (Do et al., 2008), MXSCARNA version 2 (Tabes et al., 2008) and the tools of the common secondary structure prediction from multiple alignments of RNA sequences: PETfold v1.0 (Seemann et al., 2008), RNAalifold version 1.8.1 (Bernhart et al., 2008) which was recently updated, and CentroidFold version 0.0.4 (Hamada et al., 2008) with the CONTRAfold model which implemented the averaged \( \gamma \)-centroid estimators. (M3) Simultaneously align and fold \( [x] \cup D \) and map the common secondary structure on the structural multiple alignment to the target sequence \( x \). We used RAf v1.00 (Do et al., 2008) in this method. Note that this method can calculate the most likely structural alignment and its common secondary structure, sacrificing extremely huge computational time compared with M1, M2 and the proposed method.

### 3.3 Comparison within proposed methods

In this section, we mainly evaluated the performance of the proposed estimators with the various probabilistic models of \( p^{\text{M1}} \) and \( p^{\text{M2}} \), and the various parameters in the proposed method. We used the probability distribution \( p^{\text{M1}}(\cdot) \) on \( \mathcal{A}(x, x') \) for two RNA sequences \( x \) and \( x' \) in the McCaskill model (McCaskill, 1990) and the CONTRAfold model (Do et al., 2006a), and also used the probability distribution \( p^{\text{M2}}(\cdot) \) on \( \mathcal{A}(x, x') \) for two RNA sequences \( x \) and \( x' \) in the ProbCons model (Do et al., 2005) and the ProbAlign model (Roshan and Livesey, 2006). Note that the fixed alignment is not required in the proposed method, while the method (M2) requires fixed alignments.

First, we aimed at determining the influence of the parameters \( \alpha \) (which is used only in approx3-1) and \( \delta \) in the proposed estimators on the dataset1. Figure 6a and Supplementary Figure S1 show the results of various \( \alpha \) parameter when we set \( \delta = 0 \). These show the suitability of \( \alpha \) parameter in the approx3-1 for around 0.1, which is equal to \( \alpha = 1/(|D|+1) \) where \( |D| = 9 \) in this experiment since each test data in dataset1 contains nine homologous sequences. This suggests that the same weight between the target RNA sequence \( x \) and each sequence in \( D \) provided good prediction performance, and

![Fig. 6. Performance of the different parameters for approx3-1. (a) We tested various values of \( \alpha \) for \( \delta = 0 \). (b) We tested various values of \( \delta \) for \( \alpha = 0.1 \). The performance of the other combinations of probability distributions is shown in the Supplementary Material (Figs S1–S3).](https://academic.oup.com/bioinformatics/article-abstract/25/12/i330/192269)
it seems to be a natural result. We also investigated the influence of the parameter $\delta$ described in Section 2.2.3. As shown in the Figure 6b (also see Supplementary Figs S2 and S3) in the case that $\delta < 0.01$, the performance is almost the same as that of $\delta = 0$. On the other hand, the total calculation times of the proposed estimator $\delta = 0.01$ are much faster than that of $\delta = 0$ (Table 2). Therefore, we decided to use $\alpha = 1/(|D| + 1) = 0.1$ and $\delta = 0.01$ in next experiments.

Second, we conducted an experiment for comparing appro3-1 and appro3-2 as described in Section 2.2.3. As shown in Figure 7a, appro3-1 outperformed appro3-2 under the same conditions (i.e. probability distributions $p^{(a)}$ and $p^{(b)}$). We can observe that the best combination of the proposed estimators and the probability distributions is appro3-1 with the CONTRAfold model and the ProbCons model. Moreover, in order to examine the difference in performance of appro2-1 as described in Section 2.2.2 and appro3-1 as described in Section 2.2.3, we implemented the variant of the forward–backward algorithm for computing

Table 2. Total calculation time in seconds (predicting secondary structures for 17 $\gamma$ values of 850 RNA sequences) of the estimator appro3-1 for each $\delta$ and each combination of the probability distributions

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>pa-ct</th>
<th>pa-mc</th>
<th>pc-ct</th>
<th>pc-mc</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>276.609</td>
<td>290.579</td>
<td>274.758</td>
<td>290.174</td>
</tr>
<tr>
<td>0.0001</td>
<td>16.471</td>
<td>51.073</td>
<td>16.393</td>
<td>50.09</td>
</tr>
<tr>
<td>0.001</td>
<td>79.62</td>
<td>320.2</td>
<td>79.97</td>
<td>312.4</td>
</tr>
<tr>
<td>0.01</td>
<td>3836</td>
<td>2130</td>
<td>3791</td>
<td>2065</td>
</tr>
<tr>
<td>0.1</td>
<td>2552</td>
<td>1600</td>
<td>2502</td>
<td>1541</td>
</tr>
</tbody>
</table>

We set $\alpha = (|D| + 1)^{-1}$ in this experiment. `$x-y$' means appro3-1 with $x$ of $p^{(a)}$ and $y$ of $p^{(b)}$. $\alpha$, pa, pc, mc and ct denote the ProbAlign model (Roshan and Livesay, 2006), the CONTRAfold model (Do et al., 2005), the McCaskill model (McCaskill, 1990) and the CONTRAfold model (Do et al., 2006a), respectively.

3.4 Comparison with the other methods

In this section, we compared the proposed estimator with the other existing methods described in Section 3.2. For our method, we chose `appro3-1 with the CONTRAfold model and the ProbCons model for $\alpha = 1/(|D| + 1)$ and $\delta = 0.01$' ($|D|$ is the number of homologous sequences), which achieved the best performance in the above experiments in several combinations of estimators, $p^{(a)}$ and $p^{(b)}$. This combination is denoted with the ‘proposed’ method below. First, we conducted an experiment on the dataset1 (see Section 3.1.1), which was also used in the previous section, and the result is shown in Figure S4 in the Supplementary Material. This result shows that the proposed method is better or comparable performance compared with the method (M1), (M2) and (M3). In this experiment, however, someone may be concerned about the over-fitting of the parameters $\alpha$ and $\delta$ in the proposed method (although the used parameters are very natural) because we determined the parameters on the dataset1 or simplification of the dataset1. Hence, we also conducted other experiments on the dataset2 and the dataset3. Figure 8 shows the result of the experiment on the dataset2. Remarkably, the approaches of the common secondary structure predictions (M2) with ProbCons and MXSCARNA alignments gave worse performance than the single secondary structure predictions (M1). This shows that quality of alignment strongly influences the performance of the prediction in the method (M2). We think that this is because alignment errors influence the performance and common secondary structures are not always good for a predicted secondary structure of the target.
sequence. Note that the dataset2 contains a number of sequences in the family such as Signal Recognition Particle RNA (SRP RNA), Transfer Messenger RNA (tmRNA) and RNaseP RNA, which contain diverse sequences from a number of organisms (Andronescu et al., 2008).

We also conducted an experiment on the dataset3 (See Section 3.1.2), which contains longer RNA sequences than the dataset2. Due to the computational cost, we conducted the experiment for the proposed method, the methods (M1) and (M2) with ProbCons (Do et al., 2005). The result shown in Figure 9 indicates that the proposed method outperforms the other methods such as probcons-centroid (common secondary structure prediction after aligning target and homologous sequences), simfold (secondary structure prediction from target sequences) and so forth. Table 3 also shows that the proposed method can predict secondary structures within practical calculation time even for long sequences. The experiments on the dataset1 and the dataset2 demonstrate that the methods using RAF (Do et al., 2008) are comparable with the proposed method, but these are obviously much slower than the proposed method (Table 3). It should be also emphasized that the proposed method is much better than ‘probcons-centroid’ [that predicts a common secondary structure by using averaged γ-centroid estimators (Hamada et al., 2008) after aligning the target sequence and homologous sequences by ProbCons (Do et al., 2005)], although these two methods employ the similar features (that is, both methods use base-pairing probabilities of the target sequence and homologous sequences given by the CONTRAfold model). The proposed method uses alignment probabilities between the target sequence and each homologous sequence given by the ProbCons model, whereas ‘probcons-centroid’ use the fixed alignment given by ProbCons.

Fig. 8. The result on the dataset2 (see Section 3.1.2). ‘Proposed’ means the approx3-1 with the CONTRAfold model and the ProbCons model of $\alpha = 1/(1 + |D|)$ and $\beta = 0.01$. ‘X-Y (M2)’ means the method that predicts a common secondary structure by X after aligning $\{|x|\cup D\}$ by Y, and then maps the common secondary structure to the target sequence $x$. ‘raf (M3)’ means the method that simultaneously aligns and folds $\{|x|\cup D\}$ by using RAF (Do et al., 2008), and then maps the common secondary structure to the target sequence $x$. ‘centroid’ indicates the γ-centroid estimators (Hamada et al., 2008) for the method (M1) and the averaged γ-centroid estimators (Hamada et al., 2008) for the method (M2), respectively. ‘alifold’ means RNAalifold (Bernhart et al., 2008).

Fig. 9. The result on the dataset3 (Section 3.1.2). See also the caption of Figure 8.
Table 3. Total calculation time in seconds for each dataset and each prediction method

<table>
<thead>
<tr>
<th>Methods</th>
<th>dataset1</th>
<th>dataset2</th>
<th>dataset3</th>
</tr>
</thead>
<tbody>
<tr>
<td>proposed</td>
<td>3781</td>
<td>42767</td>
<td>299941</td>
</tr>
<tr>
<td>centroid (M1)</td>
<td>339</td>
<td>2712</td>
<td>26832</td>
</tr>
<tr>
<td>raf (M1)</td>
<td>97</td>
<td>291</td>
<td>648</td>
</tr>
<tr>
<td>simfold (M1)</td>
<td>228</td>
<td>1522</td>
<td>5264</td>
</tr>
<tr>
<td>mxscarna-centroid (M2)</td>
<td>12033</td>
<td>170427</td>
<td>—</td>
</tr>
<tr>
<td>mxscarna-alifold (M2)</td>
<td>9845</td>
<td>137373</td>
<td>—</td>
</tr>
<tr>
<td>mxscarna-petfold (M2)</td>
<td>15062</td>
<td>206548</td>
<td>—</td>
</tr>
<tr>
<td>probcons-centroid (M2)</td>
<td>3097</td>
<td>35359</td>
<td>219538</td>
</tr>
<tr>
<td>probcons-alifold (M2)</td>
<td>1001</td>
<td>8069</td>
<td>33012</td>
</tr>
<tr>
<td>probcons-petfold (M2)</td>
<td>7455</td>
<td>93128</td>
<td>402333</td>
</tr>
<tr>
<td>raf-centroid (M2)</td>
<td>68684</td>
<td>1052753</td>
<td>—</td>
</tr>
<tr>
<td>raf-alifold (M2)</td>
<td>66370</td>
<td>10494276</td>
<td>—</td>
</tr>
<tr>
<td>raf-petfold (M2)</td>
<td>71818</td>
<td>10570216</td>
<td>—</td>
</tr>
<tr>
<td>raf (M3)</td>
<td>66180</td>
<td>10493751</td>
<td>—</td>
</tr>
</tbody>
</table>

Fig. 10. An example of predicted secondary structures of tRNA (ID:SPR_00633) in the RNA STRAND database (Andronescu et al., 2008). We used eight homologous sequences containing four unusual tRNA sequences (SPR_00397,00629,00832,00938) whose secondary structures lack the D-domain stem-loop.

Figure 10 and Supplementary Figure S5 show a typical example which supports the advantage of the proposed method. We used a tRNA sequence (SPR_00633) as the target sequence, which supports the advantage of the proposed method, while common secondary structure predictions after aligning the target and homologous sequences (the method (M2)) consider the only optimal alignment. The proposed method also solves Problem 1 directly, while (M2) and (M3) are indirect methods for the problem, because they predict common secondary structures instead of secondary structures of the target sequence. In the computational experiments, we confirmed that the proposed method achieved better prediction accuracy and more efficient computational time than the other methods.

Remark that a similar idea described in Section 2 leads to an estimator on the alignment problem, which relates to the probabilistic consistency transformation (PCT) (Do et al., 2005). The proposed method can be regarded as a variant of the PCT, which transforms base-pairing probabilities of each homologous sequence into those of the target sequence through alignment match probabilities between the target sequence and the homologous sequence.

RAF (Do et al., 2008) and the proposed method employ the same strategy that a probability distribution of structural pairwise alignment $p(s)$ is factorized into that of secondary structures $p(s)$ and that of pairwise alignments $p(a)$, RAF uses $p(s)$ and $p(a)$ for composing a scoring function of structural alignments and for reducing the search space of the Sankoff-style dynamic programming algorithm. Although this makes RAF to be one of the most efficient tools among several implementations of the Sankoff algorithm, RAF is still much slower than the different approaches such as (M1) and (M2) as shown in Section 3.4. The proposed method also uses $p(s)$ and $p(a)$ for composing the efficient scoring (gain) function of predicting secondary structures of target sequences by combining the information of homologous sequences, and succeeds in both keeping the quality of secondary structure predictions and reducing the computational time compared with RAF, despite the use of the same information as RAF.

In this article, we considered global pairwise alignments for computing $p(s)$. Alternatively, we can employ the local alignment models, instead of the global alignment models, which may enable to incorporate the partial homology information such as domain motifs of functional RNAs in order to further improve the secondary structure prediction of target sequences.

4 DISCUSSION AND CONCLUSIONS

In this article, we designed a novel estimator for Problem 1, based on the posterior decoding techniques. The proposed method considers all the suboptimal alignments between the target sequence and its homologous sequence, and all the suboptimal secondary structures of the target sequence and homologous sequences. This is one of the advantage of the proposed method, while common secondary structure predictions after aligning the target and homologous sequences (the method (M2)) consider the only optimal alignment. The proposed method also solves Problem 1 directly, while (M2) and (M3) are indirect methods for the problem, because they predict common secondary structures instead of secondary structures of the target sequence. In the computational experiments, we confirmed that the proposed method achieved better prediction accuracy and more efficient computational time than the other methods.

ACKNOWLEDGEMENTS

The authors thank L. E. Carvalho and C. E. Lawrence for valuable comments. The authors are also grateful to our colleagues at the Computational Biology Research Center (CBRC) for fruitful discussions.
Funding: This work was partially supported by the ‘Functional RNA Project’ funded by the New Energy and Industrial Technology Development Organization (NEDO) of Japan and a Grant-in-Aid for Scientific Research on Priority Areas ‘Comparative Genomics’ from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of Interest: none declared.

REFERENCES


Ding, Y. et al. (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA, 11, 1157–1166.


