BIOINFORMATICS

Vol. 25 ISMB 2009, pages i330–i338
doi:10.1093/bioinformatics/btp228

Predictions of RNA secondary structure by combining homologous sequence information

Michiaki Hamada[1,2,3,*], Kengo Sato[2,4], Hisanori Kiryu[2], Toutai Mituyama[2] and Kiyoshi Asai[2,5]

[1]Mizuho Information & Research Institute, Inc, 2–3 Kanda-Nishikicho, Chiyoda-ku, Tokyo 101–8443, [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2–41–6, Aomi, Koto-ku, Tokyo 135–0064, [3]Department of Computational Intelligence and System Science, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama 226–8503, [4]Japan Biological Informatics Consortium (JBIC), 2–45 Aomi, Koto-ku, Tokyo 135–8073 and [5]Graduate 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.*, 2007b).

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 $D \cup \{x\}$ (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,

[*]To whom correspondence should be addressed.

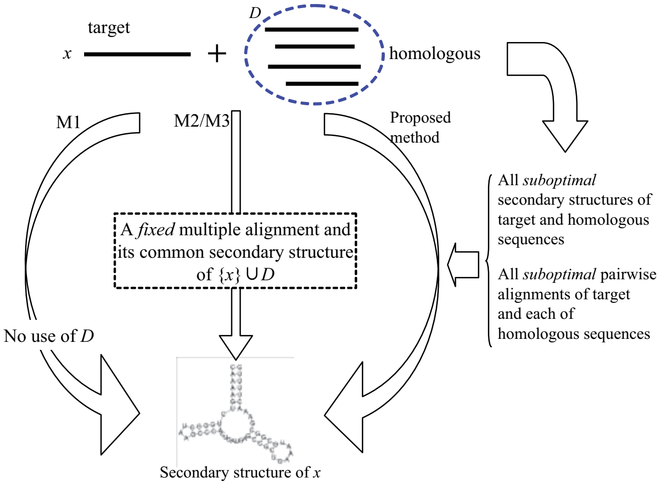

target
x
+
D
homologous
M1
M2/M3
Proposed method
All *suboptimal* secondary structures of target and homologous sequences
All *suboptimal* pairwise alignments of target and each of homologous sequences
A *fixed* multiple alignment and its common secondary structure of {x} ∪ D
No use of D
Secondary structure of x

Fig. 1. A schematic diagram of our proposed method and other existing methods (M1, M2, M3) for Problem 1. See Section 3.2 for precise descriptions of M1, M2 and M3.

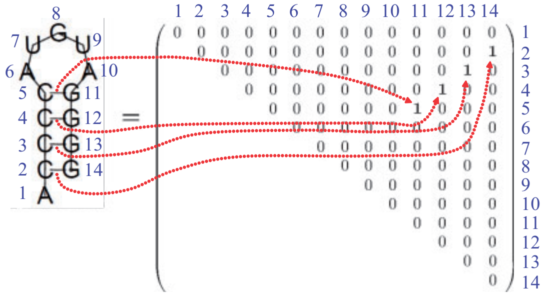
2008; Ding *et al.*, 2005; Fariselli *et al.*, 2005; Holmes and Durbin, 1998; Lunter *et al.*, 2008; Miyazawa, 1995; Paten *et al.*, 2009; Wong *et al.*, 2008). In secondary structure prediction of RNA, posterior decoding methods are used in the maximum expected accuracy (MEA) estimator of CONTRAfold (Do *et al.*, 2006a), the centroid estimator (Carvalho and Lawrence, 2008; Ding *et al.*, 2005) and the γ-centroid estimator of CentroidFold (Hamada *et al.*, 2008). We demonstrate that the prediction accuracy of the proposed method outperforms that of conventional secondary structure prediction methods.

2 METHODS

2.1 Discrete spaces and probability distributions

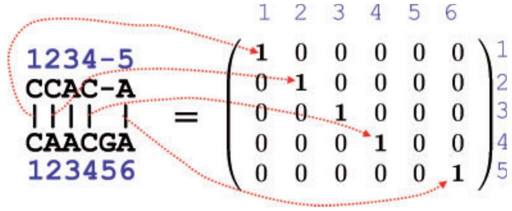
In this article, the following discrete spaces and probability distributions on those spaces play important roles in order to define proposed estimators.

2.1.1 A space of secondary structures: $\mathcal{S}(x)$ A secondary structure of an RNA sequence x is represented as an upper triangular binary matrix $\theta=\{\theta_{ij}\}_{i<j}$, where the (i,j) element θ_{ij} of the matrix θ is equal to 1 if x_i and x_j form a base pair and to 0 if x_i and x_j do not form a base pair (Fig. 2). A consistent secondary structure represented by θ should satisfy the following constraints: (i) $\sum_{i<j}\theta_{ij}+\sum_{j<i}\theta_{ji}\leq 1$ for any j (each position in a sequence is allowed to form a base pair with one other base at most), and (ii) $\theta_{ij}+\theta_{kl}\leq 1$ for any $i<k<j<l$ (the formation of any two base pairs which results in a pseudo-knot is not allowed). We denote with $\mathcal{S}(x)$ the space of all secondary structures θ of an RNA sequence x. Note that to predict the secondary structure of x is equivalent to predicting a point θ in $\mathcal{S}(x)$. We obtain a probability distribution $p^{(s)}(\cdot|x)$ on $\mathcal{S}(x)$ by using the McCaskill model (McCaskill, 1990), the CONTRAfold model (Do *et al.*, 2006a), the Simfold model (Andronescu *et al.*, 2007) and the SCFG model (Dowell and Eddy, 2004). By using the distributions, a base-pairing probability $p_{ij}^{(s,x)}$, which is the probability with which x_i and x_j form a base pair, is computed as $p_{ij}^{(s,x)}=p^{(s)}(\theta_{ij}=1|x)=\sum_{\theta\in\mathcal{S}(x)}I(\theta_{ij}=1)p^{(s)}(\theta|x)$, where $I(\cdot)$ is the indicator function, which takes a value of 1 or 0 depending on whether the condition constituting its argument is true or false. We refer to $\{p_{ij}^{(s,x)}\}_{i<j}$ as a base pairing probability matrix, which can be computed by using the inside–outside algorithm (e.g. McCaskill, 1990), whose time complexity is $O(|x|^3)$ where $|x|$ is the length of x.



$$\begin{array}{cccccccccccccc|c}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 \\
 & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 3 \\
 & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 4 \\
 & & & & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 5 \\
 & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\
 & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7 \\
 & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 \\
 & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 9 \\
 & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 10 \\
 & & & & & & & & & & 0 & 0 & 0 & 0 & 11 \\
 & & & & & & & & & & & 0 & 0 & 0 & 12 \\
 & & & & & & & & & & & & 0 & 0 & 13 \\
 & & & & & & & & & & & & & 0 & 14
\end{array}$$

Fig. 2. Representation of a secondary structure of an RNA sequence.



$$\begin{array}{cccccc|c}
1 & 2 & 3 & 4 & 5 & 6 & \\
1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 2 \\
0 & 0 & 1 & 0 & 0 & 0 & 3 \\
0 & 0 & 0 & 1 & 0 & 0 & 4 \\
0 & 0 & 0 & 0 & 0 & 1 & 5
\end{array}$$

Fig. 3. Representation of a pairwise alignment of two RNA sequences.

2.1.2 A space of pairwise alignments: $\mathcal{A}(x,x')$ Similarly to the secondary structure of an RNA sequence, a pairwise alignment of two sequences x and x' (e.g. DNA sequences, RNA sequences or amino acid sequences) is represented as a binary matrix $\theta=\{\theta_{ik}\}_{i,k}$, where $\theta_{ik}=1$ if x_i is aligned with x'_k and $\theta_{ik}=0$ if x_i is not aligned with x'_k (Fig. 3). A consistent alignment represented by θ should satisfy the following constraints: (i) $\sum_k\theta_{ik}\leq 1$ for any i, (ii) $\sum_i\theta_{ik}\leq 1$ for any k, (iii) $\theta_{ik}+\theta_{jl}\leq 1$ for any $i<j$ and $l<k$. In this article, the space of all pairwise alignments of x and x' is denoted by $\mathcal{A}(x,x')$. We can obtain the probability distribution $p^{(a)}(\theta|x,x')$ on $\mathcal{A}(x,x')$ by using the PHMM model (Durbin *et al.*, 1998), the Miyazawa model (Miyazawa, 1995), the ProbAlign model (Roshan and Livesay, 2006), the CONTRAlign model (Do *et al.*, 2006b). By using the distributions, the alignment match probability $p_{ik}^{(a,x,x')}$, which is the probability that x_i is aligned with x'_k, is computed as $p_{ik}^{(a,x,x')}=p^{(a)}(\theta_{ik}=1|x,x')=\sum_{\theta\in\mathcal{A}(x,x')}I(\theta_{ik}=1)p(\theta|x,x')$. We refer to $\{p_{ik}^{(a,x,x')}\}_{i,k}$ as an alignment match probability matrix, which can be computed by using the forward–backward algorithm (Durbin *et al.*, 1998), whose time complexity is $O(|x||x'|)$. Moreover, the joint probability $p_{ik,jl}^{(a,x,x')}$, which is the probability that x_i is aligned with x'_k and x_j is aligned with x'_l, is also computed as $p_{ik,jl}^{(a,x,x')}=p^{(a)}(\theta_{ik}=1,\theta_{jl}=1|x,x')=\sum_{\theta\in\mathcal{A}(x,x')}I(\theta_{ik}=1)I(\theta_{jl}=1)p^{(a)}(\theta|x,x')$. Note that $\{p_{ik,jl}^{(a,x,x')}\}_{i<j,k<l}$ can be computed by using a variant of the forward–backward algorithm whose time complexity is equal to $O(|x|^2|x'|^2)$.

2.1.3 A space of structural pairwise alignments of RNA sequences: $\mathcal{SA}(x,x')$ A structural alignment (e.g. Kiryu *et al.*, 2007a; Sankoff, 1985) of two RNA sequences x and x' is represented as $\theta=\{\theta_{ijkl}^{(p)}\}_{i<j,k<l}\times\{\theta_{uv}^{(l)}\}_{u,v}$ where $\theta_{ijkl}^{(p)}=1$ if a base pair (x_i,x_j) in x is aligned with a base pair (x'_k,x'_l) in x' and $\theta_{uv}^{(l)}=1$ if x_u in a loop region of x is aligned with x'_v in a loop region of x' (Fig. 4). In accordance with a consistent structural alignment, each element in θ cannot take an arbitrary value in $\{0,1\}$, where θ satisfies several conditions. Moreover, a space of structural alignments of two RNA sequences x and x' is denoted as $\mathcal{SA}(x,x')$. We obtain a probability distribution $p^{(sa)}(\theta|x,x')$ by the pair SCFG model, the Sankoff model (Sankoff, 1985). We also denote a probability that a base pair (x_i,x_j) in x is aligned with a base pair (x'_k,x'_l) in x' as $p_{ijkl}^{(sa,p,x,x')}=p^{(sa)}(\theta_{ijkl}^{(p)}=1|x,x')$. Note that $\{p_{ijkl}^{(sa,p,x,x')}\}_{i<j,k<l}$ can be

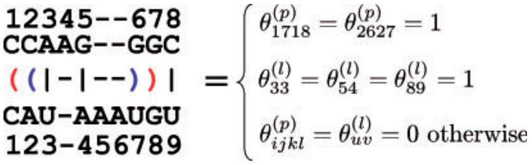

12345--678
CCAAG--GGC
((|-|--))|
CAU-AAAUGU
123-456789
=
$\theta^{(p)}_{1718} = \theta^{(p)}_{2627} = 1$
$\theta^{(l)}_{33} = \theta^{(l)}_{54} = \theta^{(l)}_{89} = 1$
$\theta^{(p)}_{ijkl} = \theta^{(l)}_{uv} = 0$ otherwise

Fig. 4. Representation of a structural 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 $O(|x|^3|x'|^3)$.

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 Y is a space from which we would like to obtain a prediction, referred to as a *predictive space*, Θ 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 Θ given a dataset d, and $G(\theta,y)$ is a *gain function* between Θ and Y, then an estimator on the predictive space is defined as

$$\hat{y} = \arg\max_{y\in Y} E_{\theta|d}[G(\theta,y)] = \arg\max_{y\in Y} \sum_{\theta\in\Theta} G(\theta,y)p(\theta|d) \tag{1}$$

(cf. Ding *et al.*, 2005; Do *et al.*, 2006a; 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 $\alpha_1 TP + \alpha_2 TN - \alpha_3 FN - \alpha_4 FN$ ($\alpha_n > 0$, $n = 1,2,3,4$) with respect to a probability distribution on $\mathcal{S}(x)$, 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 Y is $\mathcal{S}(x)$. Therefore, it is necessary to define the parameter space Θ and the probability distribution $p(\theta|x)$ on the parameter space Θ. In the following sections, we introduce several estimators: an estimator with less heuristics, and its approximations with some heuristics.

2.2.1 Estimator 1 Our assumption in Problem 1 is that x and every $x' \in D$ share a common secondary structure. The common secondary structure is naturally represented as a structural alignment of x and x' (See Section 2.1.3). Therefore, we set the parameter space as $\Theta^{(1)} = \prod_{x'\in D} \mathcal{SA}(x,x')$ and the probability distribution on $\Theta^{(1)}$ as $p^{(1)}(\theta|x,D) = \prod_{x'\in D} p^{(sa)}(\theta^{xx'}|x,x')$ for $\theta = \prod_{x'\in D} \theta^{xx'} \in \prod_{x'\in D} \mathcal{SA}(x,x')$, where $\theta^{xx'}$ is a structural alignment of x and x' (i.e. $\theta^{xx'} \in \mathcal{SA}(x,x')$). Moreover, the gain function is defined by

$$\begin{aligned} G^{(1)}(\theta,y) &= \sum_{i<j} G^{(1)}_{ij}(\theta,y_{ij}) \\ G^{(1)}_{ij}(\theta,y_{ij}) &= \gamma I(y_{ij}=1)\cdot \tfrac{1}{|D|}\sum_{x'\in D}\sum_{k<l} I(\theta^{xx'}_{ijkl}=1) \\ &\quad + I(y_{ij}=0)\left(1 - \tfrac{1}{|D|}\sum_{x'\in D}\sum_{k<l} I(\theta^{xx'}_{ijkl}=1)\right) \end{aligned}$$

for a prediction $y \in \mathcal{S}(x)$, where $|D|$ denotes with the number of sequences in D and $\gamma > 0$ 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 gain function, $\frac{1}{|D|}\sum_{x'\in D}\sum_{k<l} I(\theta^{xx'}_{ijkl}=1)$ is equal to the averaged number of a pair (x_i,x_j) which forms a base pair in the set of structural alignments $\{\theta^{xx'}\}_{x'\in D}$ and $1 - \frac{1}{|D|}\sum_{x'\in D}\sum_{k<l} I(\theta^{xx'}_{ijkl}=1)$ is equal to the averaged number of a pair (x_i,x_j) which does not form a base pair in the set. Hence, the more base pairs (x_i,x_j) there exist in the set of structural alignments, the more gain in the gain function is given for the prediction $y_{ij} = 1$. Then, we define the estimator which maximizes the expectation of the gain function $G^{(1)}(\theta,y)$ on the probability distribution $p^{(1)}(\theta|x,D)$ as

$$\hat{y} = \arg\max_{y\in\mathcal{S}(x)} \sum_{\theta\in\Theta^{(1)}} G^{(1)}(\theta,y)p^{(1)}(\theta|x,D). \tag{2}$$

It is clear that the estimator is equivalent to the γ-centroid estimator (Hamada *et al.*, 2008) on $\mathcal{S}(x)$ when the probability distribution is defined by $p(\theta|x,D) = \frac{1}{|D|}\sum_{x'\in D}\sum_{\theta'\in\Phi^{-1}(\theta)} p^{(sa)}(\theta'|x,x')$, where Φ is the natural map from a structural alignment $\theta' \in \mathcal{SA}(x,x')$ to the secondary structure $\theta \in \mathcal{S}(x)$. The optimal secondary structure is computed by a Nussinov-type dynamic programming (Nussinov *et al.*, 1978) as follows:

$$M_{i,j} = \max\begin{cases} M_{i+1,j} \\ M_{i,j-1} \\ M_{i+1,j-1} + (\gamma+1)p^{(1)}_{ij} - 1 \\ \max_k\left[M_{i,k} + M_{k+1,j}\right] \end{cases} \tag{3}$$

where $p^{(1)}_{ij} = \frac{1}{|D|}\sum_{x'\in D}\sum_{k<l} p^{(sa,p,x,x')}_{ijkl}$. As described in Section 2.1.3, the time complexity to obtain $\left\{p^{(sa,p,x,x')}_{ijkl}\right\}_{i<j,k<l}$ is $O(|x|^3|x'|^3)$, so it requires $O\left(|D||x|^3 \max_{x'\in D}|x'|^3\right)$ time for computing $\{p^{(1)}_{ij}\}_{i<j}$. $O(|x|^3)$ time is required for the dynamic programming (3) in the case of $\{p^{(1)}_{ij}\}_{i<j}$. Therefore, the total time complexity is $O\left(|D||x|^3 \max_{x'\in D}|x'|^3\right)$. 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 $\mathcal{SA}(x,x')$ and a probability distribution on the space require huge computational cost as noted in the previous section, we replace the parameter space $\Theta^{(1)}$ and the gain function $G^{(1)}(\theta,y)$ by their approximations with some heuristics. The parameter space is approximated by the factorization of $\Theta^{(1)}$ such as $\Theta^{(2)} = \mathcal{S}(x) \times \prod_{x'\in D}\left[\mathcal{A}(x,x') \times \mathcal{S}(x')\right]$ and the corresponding probability distribution on $\Theta^{(2)}$ is defined as

$$p^{(2)}(\theta|x,D) = p^{(s)}(\theta^x|x)\prod_{x'\in D}\left[p^{(a)}\left(\theta^{xx'}|x,x'\right)p^{(s)}\left(\theta^{x'}|x'\right)\right],$$

for $\theta = \theta^x \times \prod_{x'\in D}[\theta^{xx'} \times \theta^{x'}] \in \Theta^{(2)}$, where $\theta^x \in \mathcal{S}(x)$, $\theta^{xx'} \in \mathcal{A}(x,x')$ and $\theta^{x'} \in \mathcal{S}(x')$. We also consider two kinds of approximations of $\sum_{k<l} I(\theta^{xx'}_{ijkl}=1)$ in the gain function $G^{(1)}(\theta,y)$ as follows:

$$\begin{aligned} &\sum_{k<l} I(\theta^{xx'}_{ijkl}=1) \\ &\quad\approx \alpha I(\theta^x_{ij}=1) + (1-\alpha)\sum_{k<l} I(\theta^{xx'}_{ik}=1)I(\theta^{xx'}_{jl}=1)I(\theta^{x'}_{kl}=1) \end{aligned}$$

and

$$\begin{aligned} &\sum_{k<l} I(\theta^{xx'}_{ijkl}=1) \\ &\quad\approx \sum_{k<l} I(\theta^x_{ij}=1)I(\theta^{xx'}_{ik}=1)I(\theta^{xx'}_{jl}=1)I(\theta^{x'}_{kl}=1), \end{aligned}$$

where $\alpha \in [0,1]$ is a weight parameter between the target sequence x and $x' \in D$. An intuitive description of the difference in the two approximations is shown in Figure 5. The new gain functions are represented as

$$\begin{aligned} G^{(2,1)}(\theta,y) &= \sum_{i<j} G^{(2,1)}_{ij}(\theta,y_{ij}) \\ G^{(2,1)}_{ij}(\theta,y_{ij}) &= \gamma I(y_{ij}=1)\big\{\alpha I(\theta^x_{ij}=1) \\ &\quad + \tfrac{1-\alpha}{|D|}\sum_{x'\in D}\sum_{k<l} I(\theta^{xx'}_{ik}=1)I(\theta^{xx'}_{jl}=1)I(\theta^{x'}_{kl}=1)\big\} \\ &\quad + I(y_{ij}=0)\big\{1 - \alpha I(\theta^x_{ij}=1) \\ &\quad - \tfrac{1-\alpha}{|D|}\sum_{x'\in D}\sum_{k<l} I(\theta^{xx'}_{ik}=1)I(\theta^{xx'}_{jl}=1)I(\theta^{x'}_{kl}=1)\big\} \end{aligned}$$

and

$$\begin{aligned} G^{(2,2)}(\theta,y) &= \sum_{i<j} G^{(2,2)}_{ij}(\theta,y_{ij}) \\ G^{(2,2)}_{ij}(\theta,y_{ij}) &= \gamma I(y_{ij}=1) \\ &\quad \times \tfrac{1}{|D|}\sum_{x'\in D}\sum_{k<l} I(\theta^x_{ij}=1)I(\theta^{xx'}_{ik}=1)I(\theta^{xx'}_{jl}=1)I(\theta^{x'}_{kl}=1) \\ &\quad + I(y_{ij}=0) \\ &\quad \times \left(1 - \tfrac{1}{|D|}\sum_{x'\in D}\sum_{k<l} I(\theta^x_{ij}=1)I(\theta^{xx'}_{ik}=1)I(\theta^{xx'}_{jl}=1)I(\theta^{x'}_{kl}=1)\right), \end{aligned}$$

respectively. Then, we introduce two estimators in order to maximize the expectation of $G^{(2,1)}(\theta,y)$ or $G^{(2,2)}(\theta,y)$ under the probability distribution

$p^{(2)}(\theta|x,D)$. We can obtain the secondary structure of each estimator by replacing $p_{ij}^{(1)}$ with

$$p_{ij}^{(2,1)} = \alpha p_{ij}^{(s,x)} + \frac{1-\alpha}{|D|} \sum_{x' \in D} \sum_{k<l} p_{ik,jl}^{(a,x,x')} p_{kl}^{(s,x')} \text{ and}$$

$$p_{ij}^{(2,2)} = \frac{1}{|D|} p_{ij}^{(s,x)} \sum_{x' \in D} \sum_{k<l} p_{ik,jl}^{(a,x,x')} p_{kl}^{(s,x')}$$

in Equation (3), respectively. We refer to these two estimators as 'appro2-1' and 'appro2-2', respectively. As described in Section 2.1.2, time complexity for computing $\{p_{iu,jv}^{(a,x,x')}\}$ is $O(|x|^2|x'|^2)$, and the cost is $O\big(|D||x|^2 \max_{x' \in D} |x'|^2\big)$ for obtaining $\{p_{ij}^{(2,1)}\}_{i<j}$ or $\{p_{ij}^{(2,2)}\}_{i<j}$. This computational cost 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 $p_{ik,jl}^{(a,x,x')}$ to $p_{ik}^{(a,x,x')} p_{jl}^{(a,x,x')}$, we replace $p_{ij}^{(2,1)}$ and $p_{ij}^{(2,2)}$ in the previous section by

$$p_{ij}^{(3,1)} = \alpha p_{ij}^{(s,x)} + \frac{1-\alpha}{|D|} \sum_{x' \in D} \sum_{k<l} p_{ik}^{(a,x,x')} p_{jl}^{(a,x,x')} p_{kl}^{(s,x')} \text{ and}$$

$$p_{ij}^{(3,2)} = \frac{1}{|D|} p_{ij}^{(s,x)} \sum_{x' \in D} \sum_{k<l} p_{ik}^{(a,x,x')} p_{jl}^{(a,x,x')} p_{kl}^{(s,x')},$$

respectively. Note that $p_{ij}^{(3,1)}$ and $p_{ij}^{(3,2)}$ take a value between 0 and 1, although they are *not* probabilities (more precisely, $\{p_{ij}^{(3,1)}\}_{i<j}$ cannot be obtained by any probability distribution on $\mathcal{S}(x)$). We refer to these two estimators as 'appro3-1' and 'appro3-2', respectively. If $\alpha = 1/(|D|+1)$, $\{p_{ij}^{(3,1)}\}_{i<j}$ 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 $\{p_{ij}^{(3,1)}\}_{i<j}$ or $\{p_{ij}^{(3,2)}\}_{i<j}$ still remains $O\big(|D||x|^2 \max_{x' \in D} |x'|^2\big)$. However, since the number of base pairs (x'_k, x'_l) in x' where $p_{kl}^{(s,x')} > \delta$ for a threshold δ is regarded as $O(c|x'|)$ for a constant c, $\{p_{ij}^{(3,1)}\}_{i<j}$ and $\{p_{ij}^{(3,2)}\}_{i<j}$ can be computed by $O\big(c|D||x|^2 \max_{x' \in D} |x'|\big)$ only if we sum through the base pairs (x'_k, x'_l) where $p_{kl}^{(s,x')} > \delta$ in the equation $p_{ij}^{(3,1)}$ and $p_{ij}^{(3,2)}$. By definition, we see that the proposed estimators consider all suboptimal secondary structures of the target RNA sequence (and its homologous sequences) and all suboptimal pairwise alignments between the target sequence and each of homologous sequences. The proposed estimators have three parameters, α, γ and δ: the parameter α of the estimator appro3-1 represents the weight of the target sequence relative to its homologous sequences; γ is the parameter that strikes the balance between the sensitivity and specificity of the predictions; and δ represents the probability cutoff that influences the speed and accuracy of the algorithm. In Section 3.3, we confirm the relation between these parameters and prediction performance, and decide to set $\alpha = 1/(1+|D|)$ and $\delta = 0.01$ as default parameters.

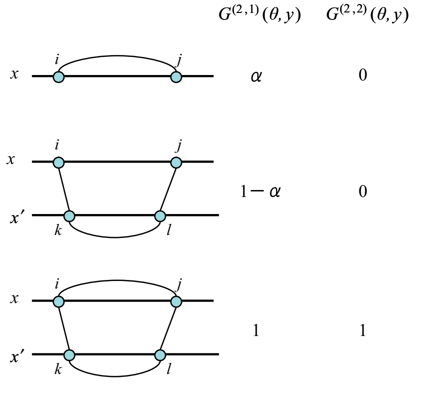


Fig. 5. The difference between two gain functions $G^{(2,1)}(\theta,y)$ and $G^{(2,2)}(\theta,y)$. Each row indicates a configuration of θ on the target sequence x and each homologous sequence x', and its positive contributions to the gain functions for $y_{ij} = 1$. For example, the middle row means that the configuration of $\theta_{ij}^{x} = 0$ and $\theta_{ik}^{xx'} = \theta_{kl}^{x'} = \theta_{jl}^{xx'} = 1$ have a positive $(1-\alpha)$ contribution for $y_{ij} = 1$ in the gain function $G^{(2,1)}$, while have no contribution in the gain function $G^{(2,2)}$.

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 $\gamma \in \{2^k : -5 \le k \le 10, k \in \mathbb{Z}\} \cup \{6\}$ for the proposed estimators, the averaged γ-centroid estimators and the γ-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

	dataset1	dataset2	dataset3
#data	850	1547	215
#homo.seqs $\lvert D\rvert$	9	2–20	2–20
Length of seqs	48–389	50–500	500–1500
Source	Kiryu *et al.* (2007b)	Andronescu *et al.* (2008)	Andronescu *et al.* (2008)

#data means the number of test data which contains a target sequence and several homologous sequences. #hom.seq means the number of homologous sequence included in each test data. See Section 3.1 for details of each dataset.

(Andronescu *et al.*, 2008), which contains manually curated RNA secondary structures of any type and organism. We selected reference secondary structures of *target* sequences out of 3704 non-redundant entries in the database, which satisfy all of the following conditions: (i) containing one molecule, (ii) not containing any ambiguous characters (i.e. any characters excluding A, U, G, C and T) (iii) not included in the family type 'Other *', 'Synthetic RNA' and 'Unknown' and (iv) whose length is more than 50 (500) nt and less than or equal to 500 (1500) nt for the dataset2 (the dataset3, respectively). We also created homologous sequences D of the target sequence by randomly selecting 2–20 sequences in the same family of 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 (Do *et al.*, 2006a), which is theoretically and experimentally superior to CONTRAfold itself. (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 (Tabei *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 γ-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 $p^{(s)}$ and $p^{(a)}$, and the various parameters in the proposed method. We used the probability distribution $p^{(s)}(\cdot)$ on $\mathcal{S}(x)$ for an RNA sequence x in the McCaskill model (McCaskill, 1990) and the CONTRAfold model (Do *et al.*, 2006a), and also used the probability distribution $p^{(a)}(\cdot)$ 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 Livesay, 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 α (which is used only in appro3-1) and δ in the proposed estimators on the dataset1. Figure 6a and Supplementary Figure S1 show the results of various α parameter when we set $\delta = 0$. These show the suitability of α parameter in the appro3-1 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

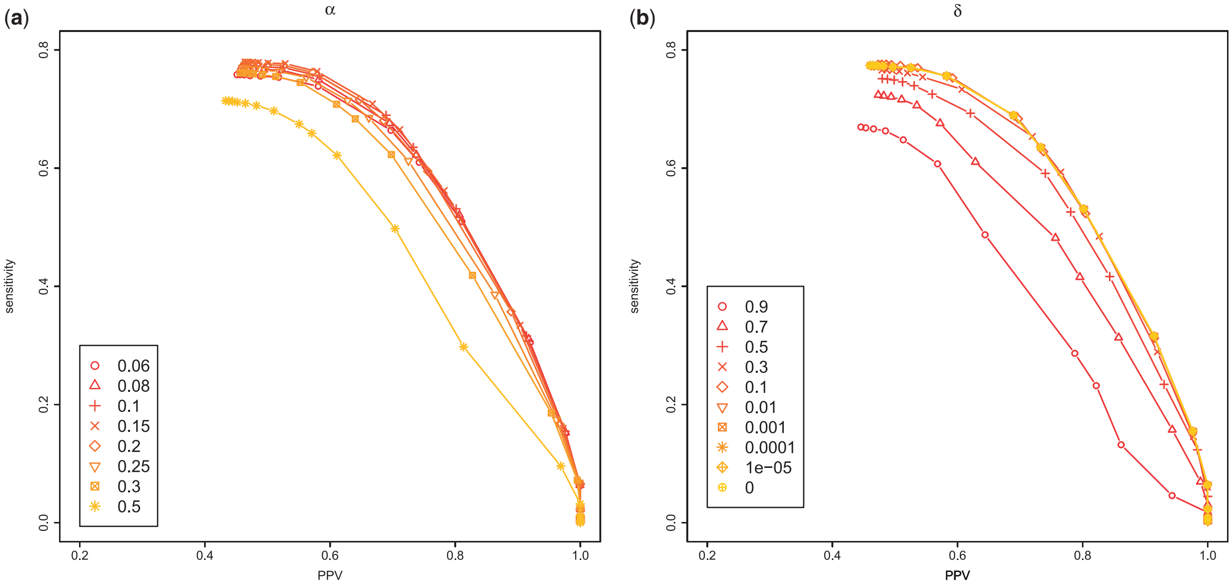


Fig. 6. Performance of the different parameters for appro3-1. (**a**) We tested various values of α for $\delta = 0$. (**b**) We tested various values of δ for $\alpha = 1/(|D|+1) = 0.1$. In both figures, we used $p^{(a)}$ in the ProbCons model and $p^{(s)}$ in the CONTRAfold model. The performance of the other combinations of probability distributions is shown in the Supplementary Material (Figs S1–S3).

it seems to be a natural result. We also investigated the influence of the parameter δ 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 $p^{(s)}$). 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 $\{p^{(a,x,x')}_{ik,jl}\}_{i<j,k<l}$ in the ProbCons model (Do *et al.*, 2005). Figure 7b shows that appro3-1 provided almost the same performance as appro2-1, indicating that the approximation used in Section 2.2.3 does not influence the performance at all.

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

δ	pa-ct	pa-mc	pc-ct	pc-mc
0	276 869	290 579	274 758	290 174
0.0001	16 471	5107	16 393	5009
0.001	7962	3202	7897	3124
0.01	3836	2130	3781	2065
0.1	2552	1600	2502	1541

We set $\alpha = 1/(|D|+1) = 0.1$ in this experiment. ‘x-y’ means appro3-1 with x of $p^{(a)}$ and y of $p^{(s)}$. pa, pc, mc and ct denote the ProbAlign model (Roshan and Livesay, 2006), the ProbCons 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 $\alpha = 1/(1+|D|)$ 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, $p^{(s)}$ and $p^{(a)}$. 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 α and δ 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*

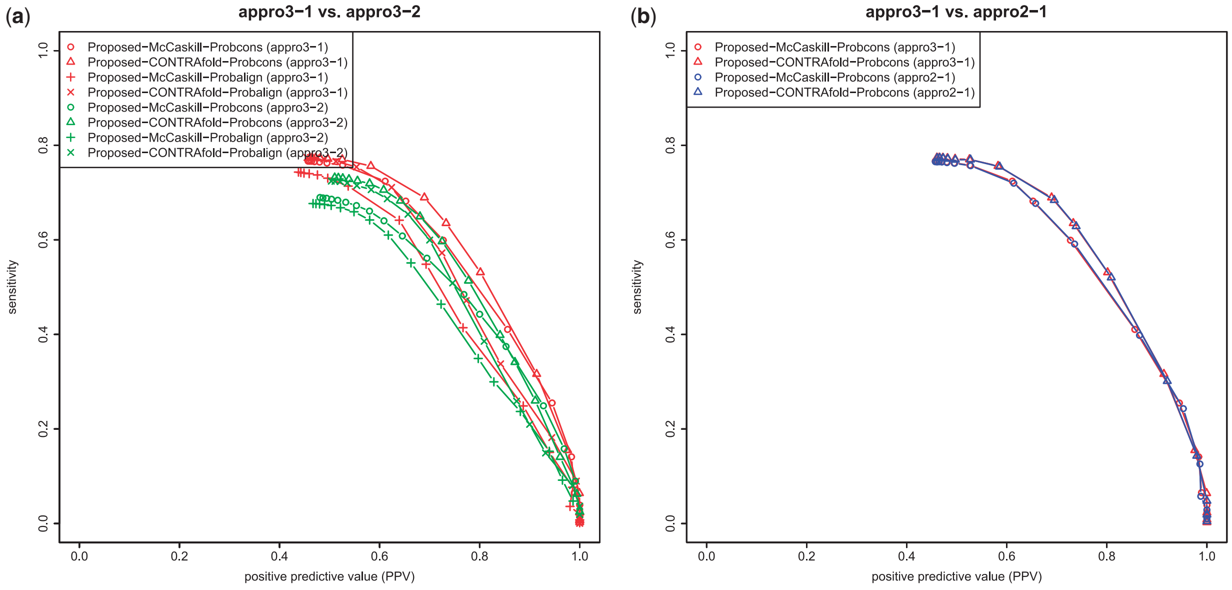


Fig. 7. (**a**) The comparison of appro3-1 and appro3-2 for various combinations of $p^{(s)}$ and $p^{(a)}$. (**b**) The comparison of appro2-1 and appro3-1 in the ProbCons model (Do *et al.*, 2005). Proposed-A-B (C) means the proposed estimator C (appro3-1, appro3-2, appro2-1 and appro2-2) with the model A [the McCaskill (McCaskill, 1990) model or the CONTRAfold model (Do *et al.*, 2006a); this is a probability distribution on $\mathcal{S}(x)$] and the model B [the ProbCons model (Do *et al.*, 2005) or the ProbAlign model (Roshan and Livesay, 2006); this is a probability distribution on $\mathcal{A}(x,x')$].

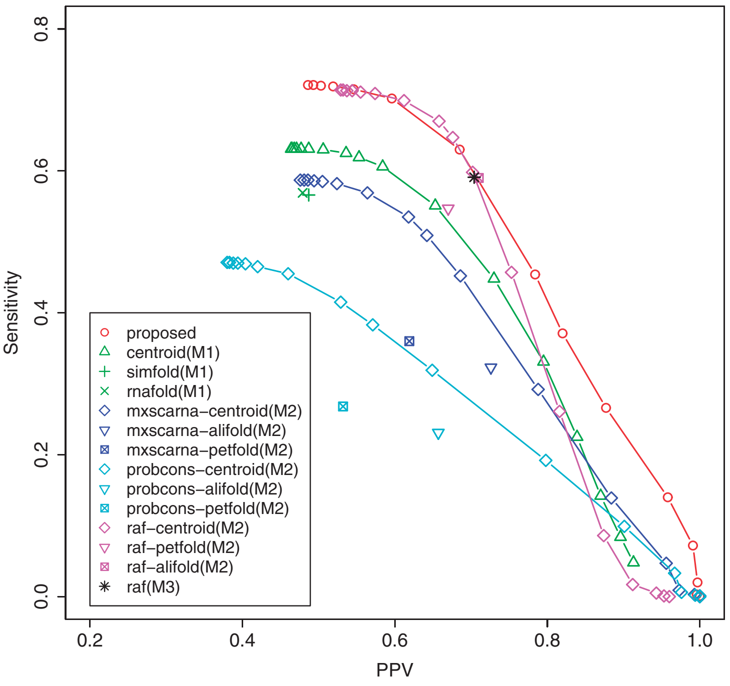


Fig. 8. The result on the dataset2 (see Section 3.1.2). 'Proposed' means the appro3-1 with the CONTRAfold model and the ProbCons model of $\alpha = 1/(1+|D|)$ and $\delta = 0.01$. 'X-Y (M2)' means the method that predicts a common secondary structure by X after aligning $\{x\} \cup D$ by Y, and then 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).

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.

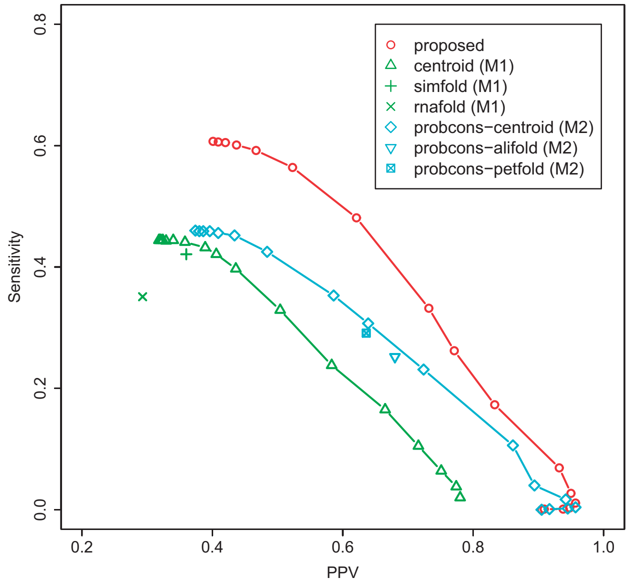
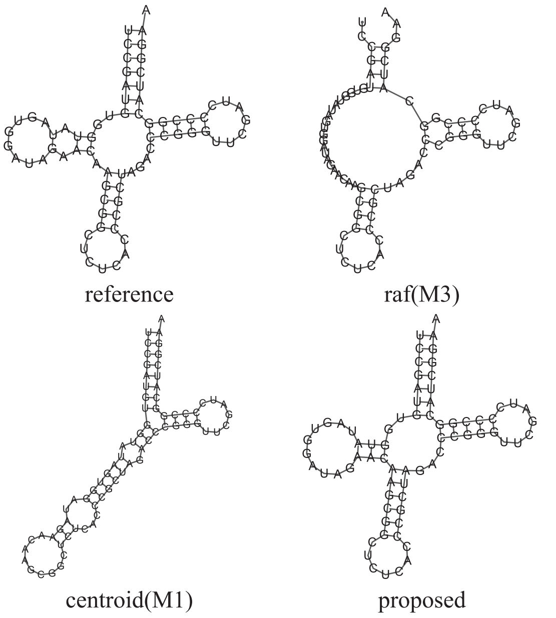


Fig. 9. The result on the dataset3 (Section 3.1.2). See also the caption of Figure 8.

Table 3. Total calculation time in seconds for each dataset and each prediction method

Methods	dataset1	dataset2	dataset3
proposed	3781	42767	299941
centroid (M1)	339	2712	26832
rnafold (M1)	97	291	648
simfold (M1)	228	1522	5264
mxscarna-centroid (M2)	12033	170428	–
mxscarna-alifold (M2)	9845	137373	–
mxscarna-petfold (M2)	15062	206548	–
probcons-centroid (M2)	3097	35359	219538
probcons-alifold (M2)	1001	8069	33012
probcons-petfold (M2)	7455	93128	402333
raf-centroid (M2)	68684	10523753	–
raf-alifold (M2)	66370	10494276	–
raf-petfold (M2)	71818	10570216	–
raf (M3)	66180	10493751	–


reference

raf(M3)

centroid(M1)

proposed

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, and eight homologous sequences containing four *unusual* tRNAs (SPR_00397,00629,00832,00938), whose secondary structures lack the D-domain stem–loop, in the RNA STRAND database (Andronescu *et al.*, 2008). These sequences are so remote that no alignment tools could produce an accurate multiple alignment of them for predicting a common secondary structure. This seems to lead to the insufficient prediction of the methods (M2) and (M3). The method (M1) also failed in predicting a secondary structure of the target sequence, whereas the proposed method could successfully predict a secondary structure of the target sequence. This result suggests that the information of the homologous sequences can improve the quality of the secondary structure prediction even if several remote homologs are included.

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.

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 $p^{(sa)}$ 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 $p^{(a)}$. 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.

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

Andronescu,M. *et al*. (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, **23**, i19–i28.

Andronescu,M. *et al*. (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.

Bernhart,S. *et al*. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Bradley,R.K. *et al*. (2008) Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics*, **24**, 2677–2683.

Carvalho,L. and Lawrence,C. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl Acad. Sci. USA*, **105**, 3209–3214.

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

Do,C. *et al*. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.

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

Do,C. *et al*. (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, **24**, i68–i76.

Do,C.B. *et al*. (2006b) Contralign: discriminative training for protein sequence alignment. In Apostolico,A. *et al*. (eds), In *Proceddings of the International Conference on Research in Computational Molecular Biology*, Vol. 3909 of *Lecture Notes in Computer Science*, Springer, pp.160–174.

Dowell,R. and Eddy,S. (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 71.

Durbin,R. *et al*. (1998) *Biological Sequence Analysis*. Cambridge University press, Cambridge.

Fariselli,P. *et al*. (2005) A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics*, **6** (Suppl. 4), S12.

Griffiths-Jones,S. *et al*. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33** (Database issue), 121–124.

Hamada,M. *et al*. (2006) Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*, **22**, 2480–2487.

Hamada,M. *et al*. (2008) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465–473.

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

Hofacker,I.L. *et al*. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.

Holmes,I. and Durbin,R. (1998) Dynamic programming alignment accuracy. *J. Comput. Biol.*, **5**, 493–504.

Kiryu,H. *et al*. (2007a) Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, **23**, 1588–1598.

Kiryu,H. *et al*. (2007b) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics*, **23**, 434–441.

Lunter,G. *et al*. (2008) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, **18**, 298–309.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Miyazawa,S. (1995) A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng.*, **8**, 999–1009.

Nussinov,R. *et al*. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math.*, **35**, 68–82.

Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.

Paten,B. *et al*. (2009) Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, **25**, 295–301.

Roshan,U. and Livesay,D. (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, **22**, 2715–2721.

Sankoff,D. (1985) Simultaneous solution of the RNA folding alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.

Seemann,S. *et al*. (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, **36**, 6355–6362.

Tabei,Y. *et al*. (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.

Wong,K.M. *et al*. (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.