BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 4 2009, pages 465–473
doi:10.1093/bioinformatics/btn601

Sequence analysis

Prediction of RNA secondary structure using generalized centroid estimators

Michiaki Hamada[1,2,3,*], Hisanori Kiryu[2], Kengo Sato[2,4], 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

Received on August 11, 2008; revised on November 5, 2008; accepted on November 14, 2008

Advance Access publication December 18, 2008

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Recent studies have shown that the methods for predicting secondary structures of RNAs on the basis of posterior decoding of the base-pairing probabilities has an advantage with respect to prediction accuracy over the conventionally utilized minimum free energy methods. However, there is room for improvement in the objective functions presented in previous studies, which are maximized in the posterior decoding with respect to the accuracy measures for secondary structures.

Results: We propose novel estimators which improve the accuracy of secondary structure prediction of RNAs. The proposed estimators maximize an objective function which is the weighted sum of the expected number of the true positives and that of the true negatives of the base pairs. The proposed estimators are also improved versions of the ones used in previous works, namely CONTRAfold for secondary structure prediction from a single RNA sequence and McCaskill-MEA for common secondary structure prediction from multiple alignments of RNA sequences. We clarify the relations between the proposed estimators and the estimators presented in previous works, and theoretically show that the previous estimators include additional unnecessary terms in the evaluation measures with respect to the accuracy. Furthermore, computational experiments confirm the theoretical analysis by indicating improvement in the empirical accuracy. The proposed estimators represent extensions of the centroid estimators proposed in Ding *et al.* and Carvalho and Lawrence, and are applicable to a wide variety of problems in bioinformatics.

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

Contact: hamada-michiaki@aist.go.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Recent research has revealed that a number of RNAs which are not translated into proteins play important roles in cells. These RNAs are called *non-coding RNAs* (*ncRNAs*), and have attracted remarkable attention (Mattick, 2005; Prabhakar *et al.*, 2006; Venkatesh *et al.*, 2006; Washietl *et al.*, 2005; Zaratiegui *et al.*, 2007). It is known that the functions of ncRNAs are often related to their respective structures. In this article, we treat two estimation (prediction) problems of structures of RNA sequences: (i) secondary structure estimation from a single RNA sequence and (ii) *common* secondary structure estimation from multiple alignments of RNA sequences.

The secondary structure prediction of RNAs is an important classical problem in bioinformatics (Durbin *et al.*, 1998; Hofacker *et al.*, 1994). The standard solution is to predict the secondary structure possessing the *minimum free energy (MFE)*, which can be calculated by using dynamic programming (DP) algorithms e.g. Mfold (Zuker and Stiegler, 1981). The MFE structure is regarded as a *maximum likelihood (ML) estimator*, which provides the highest probabilities in a probabilistic distribution over the solutions (McCaskill, 1990). However, MFE/ML structures generally have a very low probability, and in some cases are even not optimal with respect to the number of correctly predicted base pairs (Carvalho and Lawrence, 2008). Therefore, alternative estimators which consider the entire distribution over the solutions, instead of only the solution with the highest probability, have been proposed. These include the centroid estimator (Carvalho and Lawrence, 2008) used in Sfold (Ding *et al.*, 2005) and the maximum expected accuracy (MEA) estimator used in CONTRAfold (Do *et al.*, 2006). Those estimators maximize the expectation of the objective function related to the accuracy of the prediction. In this article, we propose a novel estimator which reflects the accuracy measures more directly than the MEA estimator, and show that the MEA estimator contains unnecessary terms which do not contribute to the improvement of the accuracy with respect to the predicted secondary structure (Section 2.4). The proposed estimator maximizes the expectation of $\gamma \cdot TP + TN$, where TP is the number of the true positive base pairs and TN is that of the true negatives in the

[*] To whom correspondence should be addressed.

predicted secondary structures. We refer to this estimator as a γ-centroid estimator since the estimator is equivalent to the centroid estimator proposed in Carvalho and Lawrence (2008) when $\gamma=1$. In this article, we justify the estimators from the viewpoints of the evaluation measures for the (common) secondary structure prediction and efficiency of the computation.

Based on an idea which is similar to that of the γ-centroid estimator for secondary structure prediction, we propose another estimator, referred to as an averaged γ-centroid estimator, for common secondary structure prediction from multiple alignments of RNA sequences. We show that the properties of the averaged γ-centroid estimator are similar to those of the γ-centroid estimator, e.g. the averaged γ-centroid estimator is more suitable for common secondary structure prediction than the McCaskill-MEA estimator (Kiryu *et al.*, 2007), which is an extension of the estimator proposed in Do *et al.* (2006) to the problem of common secondary structure prediction.

In addition to the theoretical analysis of the proposed estimator, computational experiments were also performed, where the proposed estimator outperformed both the MEA estimator (Do *et al.*, 2006) for secondary structure prediction and the McCaskill-MEA estimator (Kiryu *et al.*, 2007) for common secondary structure prediction.

2 METHODS

2.1 Representation of secondary structures

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. 1). A consistent secondary structure represented by θ should satisfy the following constraints: (i) $\sum_i(\theta_{ij}+\theta_{ji})\leq 1$ for any j (each position in a sequence is allowed to form 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 whose relation is a pseudo-knot is not allowed). We denote by $\mathcal{S}(x)$ the space of all secondary structures, θ, of an RNA sequence x. Since $\mathcal{S}(x)$ depends only on the length of x, we use $\mathcal{S}(\mathcal{A})$ to denote a multiple alignment $\mathcal{A}$ as the space of secondary structures of an RNA sequence whose length is equal to the length of the alignment $\mathcal{A}$ (denoted as $|\mathcal{A}|$), and we consider $\mathcal{S}(\mathcal{A})$ as a space of *common* secondary structures of a multiple alignment of RNA sequences. Using this notation, the estimation problems treated in this article are (i) predicting a secondary structure $y\in\mathcal{S}(x)$ for a given RNA sequence x (secondary structure prediction) and (ii) predicting a common secondary structure $y\in\mathcal{S}(\mathcal{A})$ for a given multiple alignment $\mathcal{A}$ of RNA sequences (common secondary structure prediction). We propose estimators for the above problems after describing the accuracy measures for the secondary structures.

2.2 Accuracy measures for secondary structures

In the process of secondary structure prediction, it is important to predict as many correct base pairs as possible while avoiding false positive (FP) base pairs. The traditional evaluation measures for secondary structure prediction, which we describe below, are suitable for this purpose.

2.2.1 Evaluation measures for secondary structure prediction We define TP, TN, FP and FN by

$$TP=\sum_{i<j}I(y_{ij}=1)I(\theta_{ij}=1), TN=\sum_{i<j}I(y_{ij}=0)I(\theta_{ij}=0),$$

$$FP=\sum_{i<j}I(y_{ij}=1)I(\theta_{ij}=0), FN=\sum_{i<j}I(y_{ij}=0)I(\theta_{ij}=1) \quad (1)$$

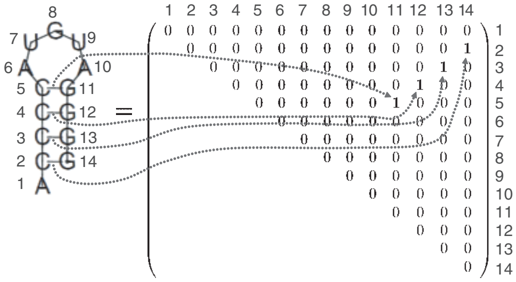
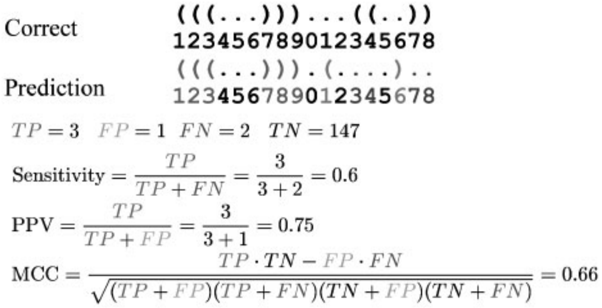


Fig. 1. A binary matrix representation of a secondary structure of an RNA sequence.



Correct (((...)))...((..))
123456789012345678

Prediction (((...))).(....)..
123456789012345678

$TP=3 \quad FP=1 \quad FN=2 \quad TN=147$

$$\text{Sensitivity}=\frac{TP}{TP+FN}=\frac{3}{3+2}=0.6$$

$$\text{PPV}=\frac{TP}{TP+FP}=\frac{3}{3+1}=0.75$$

$$\text{MCC}=\frac{TP\cdot TN-FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}=0.66$$

Fig. 2. An example of the calculation of sensitivity, PPV and MCC.

for two secondary structures $\theta\in\mathcal{S}(x)$ and $y\in\mathcal{S}(x)$. $I(\cdot)$ is the indicator function, which takes a value of 1 or 0 depending on whether the condition constituting its argument is true or false. If y is a *predicted* secondary structure and θ is a *correct* (*reference*) structure, then TP is equal to the number of correctly predicted base pairs, TN is equal to the number of base pairs which were correctly predicted as non-matching, FN is equal to the number of base pairs in the correct structure which were not predicted [false negatives (FNs)] and FP is equal to the number of incorrectly predicted base pairs. Then, standard and traditional evaluation measures, namely the sensitivity, the positive predictive value (PPV) and Matthew's correlation coefficient (MCC), are defined as follows: Sensitivity$=\frac{TP}{TP+FN}$, PPV$=\frac{TP}{TP+FP}$ and

$$\text{MCC}=\frac{TP\cdot TN-FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

When the sensitivity is equal to 1, the predicted structure contains all of the correct base pairs (although it might also contain false base pairs) and when the PPV is equal to 1, the predicted secondary structure contains only the correct base pairs. Therefore, a trade-off relation exists between the sensitivity and PPV, and perfect prediction is achieved if both the sensitivity and PPV are equal to 1. On the other hand, MCC is an evaluation measure related to the balance between the sensitivity and PPV. Note that the only correctly predicted base pairs (and correctly predicted non-base pairs) in the predicted secondary structure can contribute to the improvement of the measures. We show an illustrated example of the measures in Figure 2.

2.2.2 Evaluation measures for common secondary structure prediction When the correct secondary structure of every RNA sequence in the input alignment and a predicted common secondary structure of the alignment are given, we evaluate the accuracy of the predicted common secondary structures as follows: (i) The predicted common secondary structure is mapped onto each sequence in the input alignment (Fig. 3). All gaps in each sequence and the corresponding base pairs in the mapped secondary structure are removed in order to maintain the consistency of the secondary structures. (ii) TP, TN, FP and FN (Section 2.2.1) are calculated for

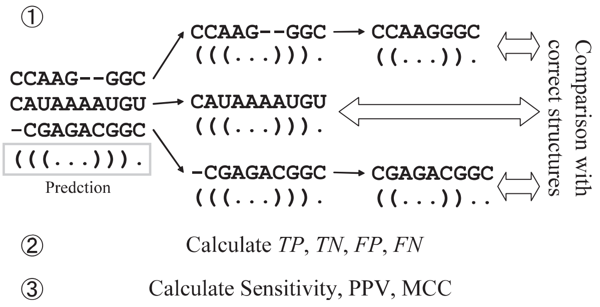


Fig. 3. An example of evaluation for the predicted common secondary structure. We assume that the correct secondary structure of every sequence in the alignment is given.

each mapped secondary structure. (iii) The sensitivity, PPV and MCC (Section 2.2.1) are calculated for the sum of *TP*, *TN*, *FP* and *FN* for all RNA sequences in the alignment. These measures are also suitable for evaluating the accuracy of common secondary structure prediction since the common secondary structure should recover correct secondary structures for each sequence in the alignment. We show an illustrated example of evaluation in Figure 3.

2.3 Proposed estimators

In the previous section, we described the accuracy measures for secondary structure prediction for a single RNA sequence. Here, we propose a novel estimator based on an objective function, which is designed for improving the accuracy measures. We assume that a probability distribution $p(\theta|x)$ on $\mathcal{S}(x)$ for a given RNA sequence x, such as the McCaskill model (McCaskill, 1990), the CONTRAfold model (Do *et al.*, 2006) or the SCFG model (Dowell and Eddy, 2004) is provided. On the basis of this distribution, the proposed estimator predicts the secondary structure which maximizes the expected weighted true predictions of base pairs in the predicted structure.

2.3.1 γ-Centroid estimator for secondary structure prediction As described in the Introduction section, the usual MFE/ML solutions generally have very small probability values and many near-optimal structures have probabilities close to the maximal one due to the extremely large volume of the structure space. This indicates that the MFE/ML solution does not accurately represent the entire distribution $p(\theta|x)$ (Carvalho and Lawrence, 2008). Therefore, instead of predicting the most likely but very weakly supported solution, we consider to predict the structure that optimizes the *expected numbers* of base pairs of *TP*, *TN*, *FP* and *FN* with respect to the entire distribution $p(\theta|x)$. Let $\bar{G}(\theta,y)$ denote a general linear combination of the accuracy measure functions in Equation (1) that rewards the correctly predicted base pairs ($y_{ij}=\theta_{ij}$) and penalizes the incorrect ones ($y_{ij}\neq\theta_{ij}$),

$$\bar{G}(\theta,y)=\alpha_1 TP+\alpha_2 TN-\alpha_3 FP-\alpha_4 FN$$

where $\alpha_k>0$ ($k=1,2,3,4$) represent arbitrary constants. By using the identity $I(y_{ij}=0)+I(y_{ij}=1)=1$ in the definitions given in Equation (1), the right hand side of the above formula can be rewritten as

$$\text{r.h.s.}=C_0 G^{(c)}_\gamma(\theta,y)+C_\theta$$
$$G^{(c)}_\gamma(\theta,y)=\gamma TP+TN, \tag{2}$$

where $C_0,\gamma>0$ are constants, and C_θ is a function of θ independent of y. (see Section A.1 of the Supplementary Material for the derivation.) We then propose the estimator that predicts the structure $\hat{y}$ that maximizes the expectation value of $G^{(c)}_\gamma(\theta,y)$ with respect to $p(\theta|x)$,

$$\hat{y}=\arg\max_{y\in\mathcal{S}(x)} E_{\theta|x}[G^{(c)}_\gamma(\theta,y)]. \tag{3}$$

$$E_{\theta|x}[G^{(c)}_\gamma(\theta,y)]=\sum_{\theta\in\mathcal{S}(x)} G^{(c)}_\gamma(\theta,y)p(\theta|x). \tag{4}$$

Our method has the advantage over the MFE/ML method that the predicted base pairs are supported by a large number of near-optimal structures.

Similar to the MEA estimator proposed in Do *et al.* (2006), the γ parameter has the role of adjusting the sensitivity and PPV of the predicted secondary structures. Estimators with larger γ values produce better sensitivities (and smaller PPVs), while those with smaller γ values produce better PPVs (smaller sensitivities). We refer to the proposed estimator as a *γ-centroid estimator* as it is equivalent to the centroid estimator proposed in Carvalho and Lawrence (2008) and Ding *et al.* (2005) when $\gamma=1$. We also refer to $G^{(c)}_\gamma(\theta,y)$ as the *gain function* of γ-centroid estimators.

Equation (4) can be rewritten as follows:

$$E_{\theta|x}[G^{(c)}_\gamma(\theta,y)]=\sum_{i<j}\big[(\gamma+1)p_{ij}-1\big]I(y_{ij}=1)+C, \tag{5}$$

where $p_{ij}=p(\theta_{ij}=1|x)=\sum_{\theta\in\mathcal{S}(x)}I(\theta_{ij}=1)p(\theta|x)$ and C is a constant which does not depend on y (see Section A.2 of the Supplementary Material for the derivation.) Furthermore, $\{p_{ij}\}_{i<j}$ is the base-pairing probability matrix, which is computed efficiently by using a DP algorithm [e.g. the algorithm proposed in McCaskill (1990)]. Equation (5) indicates that the proposed estimator predicts the secondary structure which maximizes the sum of the base-pairing probabilities larger than $1/(\gamma+1)$, because the prediction $y_{ij}=0$ never leads to an inconsistent secondary structure. This means that the secondary structure which maximizes the expected value of $\gamma TP+TN$ is equivalent to the one which maximizes the sum of base-pairing probabilities larger than $1/(\gamma+1)$ in a predicted secondary structure. Moreover, such secondary structures can be efficiently computed by using a simple Nussinov-type DP algorithm (Nussinov *et al.*, 1978):

$$M_{i,j}=\max\begin{cases} M_{i+1,j} \\ M_{i,j-1} \\ M_{i+1,j-1}+(\gamma+1)p_{ij}-1 \\ \max_k\big[M_{i,k}+M_{k+1,j}\big] \end{cases}. \tag{6}$$

It should be noted that by combining the above considerations with Theorem 2 in Carvalho and Lawrence (2008), we can recover the secondary structure of the γ-centroid estimator $\hat{y}=\{\hat{y}_{ij}\}$ for $\gamma\in[0,1]$ as follows, without the need to use the DP technique presented in Equation (6)

$$\hat{y}_{ij}=\begin{cases} 1 & \text{if } p_{ij}>\frac{1}{\gamma+1} \\ 0 & \text{if } p_{ij}\le\frac{1}{\gamma+1} \end{cases} \quad \text{for } 1\le i<j\le|x|.$$

See Section A.3 in the Supplementary Material for details of the proof. This property will be useful for some situations e.g. if we pre-compute the genome-wide base-pairing probabilities by using Rfold (Kiryu *et al.*, 2008), we can quickly recover the secondary structure of the γ-centroid estimators of $\gamma\in[0,1]$ for any long sequence (although the sensitivity might not be so good). Note that the MEA estimator proposed in Do *et al.* (2006) does not have such properties (see Section B in the Supplementary Material).

2.3.2 Averaged γ-centroid estimator for common secondary structure prediction In this section, we propose novel estimators for common secondary structure prediction, whose underlying principles are similar to the ones presented in the previous section. As described in Section 2.2.2, the evaluation of predicted common secondary structures is conducted through the steps shown in Figure 3. Our proposed method is to maximize the *sum* of the expected value of $\alpha_1 TP+\alpha_2 TN-\alpha_3 FP-\alpha_4 FN$ (equivalent to

$\gamma TP+TN$) for all sequences in the input alignment, which are suitable for the evaluation (See Section A.5 of the Supplementary Material for details).

If a sequence x' in the input alignment does not contain any gaps, $p(\theta'|x')$ [the probability distribution over a space of secondary structures of x', i.e. $\mathcal{S}(x')$] can be computed using the McCaskill model (McCaskill, 1990), the CONTRAfold model (Do *et al.*, 2006) or another suitable model. On the other hand, if a sequence x in the alignment contains several gaps, the probability $p(\theta|x)$ can be defined as follows: if the secondary structure θ contains no base pairs which correspond to gaps in x, then $p(\theta|x)=p(\theta'|x')$ where θ' and x' are the secondary structure and the sequence obtained by removing all gaps in θ and x, respectively, and $p(\theta'|x')$ is a probability distribution on $\mathcal{S}(x')$ such as the McCaskill model; otherwise $p(\theta|x)=0$. We can see that $p(\theta|x)$ as defined above is a probability distribution (see Section A.4 in the Supplementary Material). Then, we define the proposed estimator as

$$\begin{aligned}\hat{y}&=\arg\max_{y\in\mathcal{S}(\mathcal{A})}\sum_{x\in\mathcal{A}}E_{\theta|x}[G_\gamma^{(c)}(\theta,y)]\\&=\arg\max_{y\in\mathcal{S}(\mathcal{A})}\sum_{x\in\mathcal{A}}\sum_{\theta\in\mathcal{S}(x)}G_\gamma^{(c)}(\theta,y)p(\theta|x),\end{aligned}$$

which maximizes the sum of the expected values of $\gamma\cdot TP+TN$ for all sequences in the input alignment [cf. Equation (4)]. Note that the estimators directly reflect the accuracy measures for common secondary structure prediction (Section 2.2.2) We can easily obtain

$$\sum_{x\in\mathcal{A}}E_{\theta|x}[G_\gamma^{(c)}(\theta,y)]=N\cdot\sum_{i<j}\big[(\gamma+1)\bar{p}_{ij}-1\big]I(y_{ij}=1)+C, \qquad (7)$$

where N is the number of sequences x in the alignment $\mathcal{A}$, C is a constant which does not depend on y and $\bar{p}_{ij}=\frac{1}{N}\sum_{x\in\mathcal{A}}p(\theta_{ij}=1|x)$. (See Section A.2 of the Supplementary Material for the derivation). The matrix $\{\bar{p}_{ij}\}_{i<j}$ is referred to as averaged base-pairing probability matrix (Kiryu *et al.*, 2007). Equation (7) also indicates that this estimator predicts the common secondary structure which maximizes the sum of averaged base-pairing probabilities which are larger than $1/(\gamma+1)$, and such common secondary structures can be computed by using a simple DP algorithm which in turn can be derived from that in Equation (6) by replacing p_{ij} with $\bar{p}_{ij}$. We refer to this estimator as *averaged γ-centroid estimator*.

Note that the proposed estimator can also be regarded as a sort of γ-centroid estimator over the space $\mathcal{S}(\mathcal{A})$, which is a space of common secondary structures of the multiple alignment $\mathcal{A}$, if we define a probability distribution over $\mathcal{S}(\mathcal{A})$ as follows:

$$p(\theta|\mathcal{A})=\frac{1}{N}\sum_{x\in\mathcal{A}}p(\theta|x),$$

where $p(\theta|x)$ is a probability distribution over $\mathcal{S}(x)$ (x might contain several gaps. See the beginning of this section). It is easily seen that $p(\theta|\mathcal{A})$ becomes a probability distribution over $\mathcal{S}(\mathcal{A})$. By using this implication together with Theorem 2 in Carvalho and Lawrence (2008), we can also prove that the averaged γ-centroid estimator for $\gamma\in[0,1]$ can be recovered without using DP by gathering all base pairs whose *averaged* base-pairing probabilities are larger than $1/(\gamma+1)$ (see Section A.7 in the Supplementary Material for details).

2.4 Relation between the proposed estimator and the MEA estimator

The γ-centroid estimator presented in the previous section is similar to the MEA estimator proposed in Do *et al.* (2006), and we discuss their relation with respect to secondary structure prediction.[1] In this context, the MEA estimator is defined by replacing the gain function $G_\gamma^{(c)}(\theta,y)$ in Equation (3) with the following function:

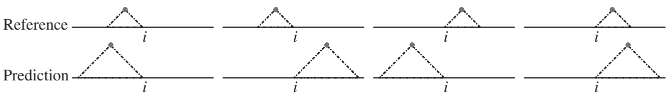


Fig. 4. Each column in the figure represents two base pairs which take the positive value of one of the last four terms in Equation (9). The red circle indicates a base pair in each RNA sequence, and no two base pairs in each column are the same.

$$G_\gamma^{(mea)}(\theta,y)=\sum_{i=1}^{|x|}\left[\gamma\sum_{j:j\neq i}I(\theta^*_{ij}=1)I(y^*_{ij}=1)+\prod_{j:j\neq i}I(\theta^*_{ij}=0)I(y^*_{ij}=0)\right]. \qquad (8)$$

where θ^* and y^* are symmetric extensions of θ and y, respectively (i.e. $\theta^*_{ij}=\theta_{ij}$ for $i<j$ and $\theta^*_{ij}=\theta_{ji}$ for $j<i$). The value of the gain function $G_\gamma^{(mea)}(\theta,y)$ is equal to the number of *bases* correctly predicted as unpaired plus γ times the number of *bases* correctly predicted as paired. This indicates that the MEA estimator maximizes the expected weighted accuracies with respect to each *base* of an input sequence, whereas the γ-centroid estimator maximizes the expected weighted accuracies with respect to each *base pair*. The following property clarifies the relation between these two estimators with respect to the respective gain functions.

PROPOSITION 1. *We obtain the relation between (2) and (8) as follows:*

$$\begin{aligned}G_\gamma^{(mea)}(\theta,y)=2G_\gamma^{(c)}(\theta,y)&+C\\&+\sum_{1\leq i\leq|x|}\sum_{\substack{j_1:j_1<i\\j_2:j_2<i\\j_1\neq j_2}}I(\theta_{j_1i}=1)I(y_{j_2i}=1)\\&+\sum_{1\leq i\leq|x|}\sum_{\substack{j_1:j_1<i\\j_2:j_2>i}}I(\theta_{j_1i}=1)I(y_{ij_2}=1)\\&+\sum_{1\leq i\leq|x|}\sum_{\substack{j_1:j_1>i\\j_2:j_2<i}}I(\theta_{ij_1}=1)I(y_{j_2i}=1) \qquad (9)\\&+\sum_{1\leq i\leq|x|}\sum_{\substack{j_1:j_1>i\\j_2:j_2>i\\j_1\neq j_2}}I(\theta_{ij_1}=1)I(y_{ij_2}=1),\end{aligned}$$

where C is a constant which does not depend on either $\theta\in\mathcal{S}(x)$ or $y\in\mathcal{S}(x)$.

The proof of this proposition is shown in Section A.1 in the Supplementary Material. This result shows that the essential difference between the gain functions of the MEA estimator and the γ-centroid estimator is in the last four terms in the right-hand side of Equation (9). These four terms contribute to $G_\gamma^{(mea)}(\theta,y)$ if the i-th base forms a base pair in both θ and y but is paired with a different base in each case (i.e. mutually incompatible base pairs; see Fig. 4). For example, if $I(\theta_{ij_1}=1)I(y_{ij_2}=1)=1$ for $j_1\neq j_2$, then we obtain $I(\theta_{ij_1}=1)I(y_{ij_1}=0)=1$ and $I(\theta_{ij_2}=0)I(y_{ij_2}=1)=1$, which implies that the base pair (i,j_1) is a FN base pair and (i,j_2) is a FP base pair when θ is a correct and y is a predicted structure. Hence, the last four terms in Equation (9) always have negative influence to the prediction. Equation (9) suggests that the additional terms influence the performance of the MEA estimator when γ is small since the last four terms do not contain γ, as well as that the two estimators are equivalent when $\gamma\to\infty$. We verify those theoretical implications later in

[1] A similar discussion of the relation between the averaged γ-centroid estimator and the McCaskill-MEA (Kiryu *et al.*, 2007) one, which is an extension of the estimator in Do *et al.* (2006) to the problem of common secondary structure prediction, is described in Section C in the Supplementary Material.

Section 3 by means of computational experiments. The expected value of Equation (8) is computed as (see Section A.2 for the details of the derivation)

$$E_{\theta|x}[G_{\gamma}^{(mea)}(\theta,y)]=\sum_{i}\Bigg[\sum_{j:j>i}\big(\gamma p_{ij}-q_i\big)I(y_{ij}=1)$$

$$+\sum_{j:j<i}\big(\gamma p_{ji}-q_i\big)I(y_{ji}=1)\Bigg]+C, \qquad (10)$$

where $p_{ij}=p(\theta_{ij}=1|x)$, $q_i=1-\sum_{j:j<i}p_{ji}-\sum_{j:j>i}p_{ij}$ and C is a constant which does not depend on $y\in\mathcal{S}(x)$. Therefore, the recursive equation for computing the secondary structure of the MEA estimator is as follows:

$$M_{i,j}=\max\begin{cases} M_{i+1,j} \\ M_{i,j-1} \\ M_{i+1,j-1}+2\gamma p_{ij}-q_i-q_j \\ \max_k\big[M_{i,k}+M_{k+1,j}\big] \end{cases}. \qquad (11)$$

Note that the recursive Equation (11) is equivalent to Equation ((11) in (Do *et al.*, 2006). Equations (6) and (11) show that the decoding methods of both estimators are identical to each other in all respects except for their scoring functions: $2\gamma p_{ij}-q_i-q_j$ for the MEA estimator and $(\gamma+1)p_{ij}-1$ for the γ-centroid estimator, respectively. Note that when $\gamma\to\infty$, both the γ-centroid estimator and the MEA estimator predict the secondary structure which maximizes the sum of the base-pair probabilities.

3 EXPERIMENTS

We carried out several experiments using three kinds of probability distribution $p(\theta|x)$ over $\mathcal{S}(x)$, namely the McCaskill model (McCaskill, 1990), the CONTRAfold model (Do *et al.*, 2006) and the Simfold model (Andronescu *et al.*, 2007). We evaluated the accuracy of the secondary structure prediction through the measures described in Section 2.2 with respect to the *correct* (*reference*) structures.

3.1 Secondary structure prediction

First, in order to evaluate the proposed estimator, we performed an experiment regarding the secondary structure prediction of an RNA sequence on the S-151Rfam dataset used in Do *et al.* (2006), and Andronescu *et al.* (2007) which contains 151 RNA sequences, each of which was taken from a different family. Figure 5 shows the PPV-sensitivity curves for the γ-centroid estimator, the MEA estimator (Do *et al.*, 2006), RNAfold (Hofacker *et al.*, 1994), which predicts the MFE structure (i.e. the ML estimator with the McCaskill model) and the ML estimator with the CONTRAfold model. We plotted the curves at $\gamma\in\{2^k:-5\le k\le 10, k\in\mathbb{Z}\}\cup\{6\}$ for the γ-centroid estimator and the MEA estimator. The total performance of the γ-centroid estimator is slightly higher than that of the MEA estimator, which is consistent with the theoretical implication that the gain function of the MEA estimator contains additional terms which lower the performance with respect to the above evaluation measures (Proposition 1). Note that the difference in performance between the MEA estimator and the γ-centroid estimator when we used the CONTRAfold model is larger than when we used the McCaskill model as the probability distribution. In order to illustrate this, in Supplementary Figure S3 we provide the histograms of the probability mass of the *true* base pairs (i.e. the base pairs present in the reference structure) and the *false* base pairs (i.e. the base pairs which are not present in the reference structure) of the dataset. Figure S3 shows that the distribution of the probability mass of the true base pairs of the McCaskill model has a strong peak around the probability of 1, while that of the CONTRAfold model exhibits a diverse distribution rather than a strong peak. Therefore, we can conclude that the diverse distribution in the CONTRAfold model makes the additional terms of the gain function of the MEA estimator larger than those of the McCaskill model. On the other hand, the distribution of the probability mass of the false base pairs of the McCaskill model also has a peak around the probability of 1, which seems to prevent the γ-centroid estimator from increasing the PPV for small γ and the sensitivity for large γ. In any case, these results provide *experimental* confirmation that the γ-centroid estimator is slightly better than the MEA estimator at predicting secondary structures of RNA sequences if the probability distribution is the same. This is in line with the theoretical analysis given in Section 2.4.

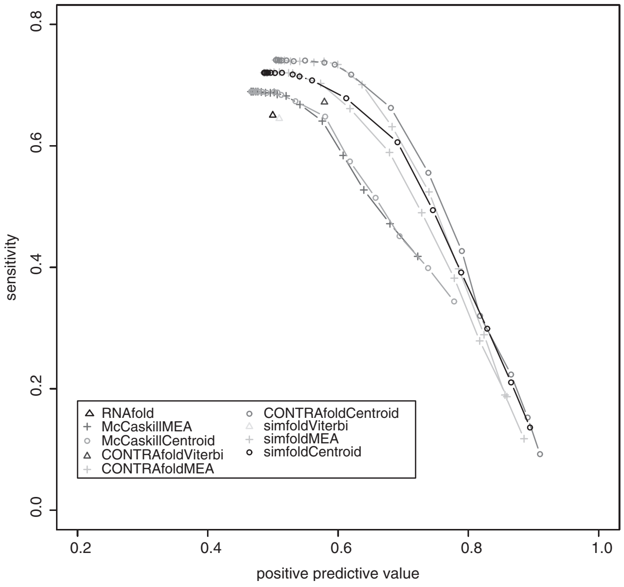


Fig. 5. Performance of secondary structure prediction of an RNA sequence. The accuracy of three estimators is shown for the following combinations: the γ-centroid estimator with the McCaskill model (McCaskill-Centroid), the γ-centroid estimator with the CONTRAfold model (CONTRAfold-Centroid), the MEA estimator with the McCaskill model (McCaskill-MEA), the MEA estimator with the McCaskill model (CONTRAfold-MEA), the ML estimator with the McCaskill model (RNAfold) and the ML estimator with the CONTRAfold model (CONTRAfold-Viterbi). We have plotted the curves at $\gamma\in\{2^k:-5\le k\le 10, k\in\mathbb{Z}\}\cup\{6\}$ for the γ-centroid estimator and the MEA estimator. The point of the largest PPV (and the lowest sensitivity) corresponds to $\gamma=2^{-5}$, and the point of the smallest PPV (and the highest sensitivity) corresponds to $\gamma=2^{10}$ in each curve. The arrow in each curve indicates the point for the MEA estimator and the γ-centroid estimator where $\gamma=1$.

We investigated the selection of the γ parameters of the γ-centroid estimator by dividing the sequences in our dataset into three subgroups: len(1), which contains sequences whose length is shorter than 80; len(2), which contains sequences whose length is between 80, inclusive and 140, exclusive; and len(3), which contains sequences whose length is 140 or above. In Supplementary Table S1, we show the MCC for various values of γ in each subgroup. The results indicate that if we need to predict the secondary structure with the highest MCC value, we should use $\gamma=1$ for the McCaskill model and $\gamma=2$ for the CONTRAfold model, respectively. This has proven that the centroid estimator proposed in Carvalho and

Table 1. MCC of the γ-centroid estimators for the secondary structure prediction

	McCaskill model				Contrafold model				Simfold model			
γ	All	Len(1)	Len(2)	Len(3)	All	Len(1)	Len(2)	Len(3)	All	Len(1)	Len(2)	Len(3)
0.03125	0.517	0.684	0.544	0.454	0.290	0.388	0.327	0.238	0.349	0.445	0.387	0.300
0.0625	0.542	0.705	0.579	0.477	0.368	0.485	0.409	0.311	0.426	0.530	0.467	0.376
0.125	0.559	0.715	0.592	0.501	0.439	0.576	0.477	0.378	0.497	0.602	0.537	0.448
0.25	0.581	0.741	0.623	0.519	0.511	0.647	0.555	0.449	0.555	0.651	0.600	0.507
0.5	0.595	0.764	0.637	0.533	0.580	0.703	0.614	0.528	0.606	0.719	0.634	0.563
1	**0.612**	**0.778**	**0.648**	**0.555**	0.640	0.761	0.682	0.587	**0.646**	**0.775**	**0.671**	0.600
2	0.599	0.766	0.641	0.540	**0.671**	0.791	**0.703**	**0.625**	0.644	0.771	0.664	**0.602**
4	0.591	0.761	0.637	0.531	0.666	**0.793**	0.695	0.622	0.629	0.766	0.641	0.589
6	0.589	0.760	0.634	0.529	0.660	0.789	0.682	0.619	0.621	0.753	0.634	0.581
8	0.587	0.758	0.631	0.526	0.652	0.783	0.672	0.612	0.616	0.750	0.629	0.577
16	0.579	0.751	0.618	0.521	0.637	0.776	0.649	0.598	0.607	0.736	0.622	0.569
32	0.576	0.743	0.614	0.519	0.626	0.765	0.640	0.586	0.601	0.723	0.615	0.565
64	0.573	0.741	0.610	0.516	0.620	0.756	0.634	0.581	0.597	0.716	0.612	0.561
128	0.571	0.738	0.607	0.514	0.616	0.747	0.630	0.577	0.595	0.713	0.610	0.559
256	0.569	0.734	0.606	0.512	0.614	0.745	0.628	0.575	0.593	0.711	0.609	0.557
512	0.567	0.731	0.604	0.511	0.612	0.742	0.627	0.574	0.592	0.709	0.608	0.556
1024	0.566	0.730	0.602	0.510	0.611	0.741	0.625	0.572	0.591	0.708	0.607	0.555

The column 'all' indicates the MCC of all sequences, 'len(1)' indicates the MCC of the set of sequences whose length is shorter than 80, 'len(2)' indicates the MCC of the set of sequences whose length is between 80, inclusive, and 140, exclusive and 'len(3)' indicates the MCC of the set of sequences whose length is 140 or above. The numbers in bold represent the largest MCC values in each column.

Lawrence (2008) and Ding *et al.* (2005) is not always optimal for the prediction of secondary structures of RNA sequences. From Table 1, we also see that the γ which provides an accurate MCC is not influenced by the length of the sequence.

3.2 Common secondary structure prediction

Next, we conducted an experiment regarding the common secondary structure prediction from multiple alignments of RNA sequences on the dataset used in Kiryu *et al.* (2007), which contains 85 reference alignments of 10 sequences taken from 17 RNA families in the Rfam database (Griffiths-Jones *et al.*, 2005). Furthermore, we also produced multiple alignments from the same sequences as the reference alignments by using four aligners: ProbCons (Do *et al.*, 2005), MAFFT (Katoh *et al.*, 2005), MXSCARNA (Tabei *et al.*, 2008) and ClustalW (Thompson *et al.*, 1994).

The panels in Supplementary Figures S4, S5, S6, S7, S8 and Figure 6, show the PPV-sensitivity curves for seven estimators of common secondary structure prediction from each multiple alignment by the reference, ProbCons, MAFFT, MXSCARNA and ClustalW, respectively, each of which is described as follows: (i) the averaged γ-centroid estimator with the McCaskill model [McCaskill-Centroid (multi)], (ii) the averaged γ-centroid estimator with the CONTRAfold model [CONTRAfold-Centroid (multi)],[2] (iii) the γ-centroid estimator with the RNAalifold model [Alipfold-Centroid (multi)], (iv) the McCaskill-MEA (Kiryu *et al.*, 2007) with the McCaskill model [McCaskill-MEA (multi)], (v) the McCaskill-MEA with the CONTRAfold model [CONTRAfold-MEA (multi)], (vi) the MEA estimator (Do *et al.*, 2006) with the RNAalifold model [Alipfold-MEA (multi)] and (vii) the ML estimator with the RNAalifold model (Hofacker *et al.*, 2002) [RNAalifold (multi); this is equivalent to the RNAalifold algorithm]. In all figures, we also plotted the PPV-sensitivity curves for the following six estimators for secondary structure prediction of a *single* sequence in each multiple alignment for the purpose of comparison: (a) the MEA estimator with the McCaskill model [McCaskill-MEA (single)], (b) the γ-centroid estimator with the McCaskill model [McCaskill-Centroid (single)], (c) the MEA estimator with the CONTRAfold model [CONTRAfold-MEA (single)], (d) the γ-centroid estimator with the CONTRAfold model [CONTRAfold-Centroid (single)], (e) the ML estimator with the McCaskill model [RNAfold (single); this is equivalent to the RNAfold algorithm], and (f) the ML estimator with the CONTRAfold model [CONTRAfold-Viterbi (single)]. The performance with respect to secondary structure prediction from multiple alignments is much higher than that of secondary structure prediction from a single sequence [i.e. (i)~(vii) > (a)~(f)[3]], which indicates that common secondary structure prediction is valuable if we can prepare the related sequences which share a common secondary structure. The results also indicate that the performance of the averaged γ-centroid estimator with a given probability distribution is always slightly higher than the McCaskill-MEA with the same probability distribution and the same input alignment [i.e. (i)≥(iv) and (ii)≥(v)], which supports our theoretical prediction that the averaged γ-centroid estimator is more suitable than the McCaskill-MEA (Section C2 in Supplementery Material). In particular, the difference in performance between the McCaskill-MEA and the averaged γ-centroid estimator is larger if the accuracy of the multiple alignment is worse since the accuracy of the alignment is thought to increase as follows: (better) Reference > MXSCARNA > Probcons $\simeq$ MAFFT > ClustalW (worse). As a speculation, misalignments tend to enlarge the last four

[2] (i) and (ii) correspond to the method proposed in this article.

[3] This notation refers to the fact that the performance of (i)~(vii) is higher than that of (a)~(f).

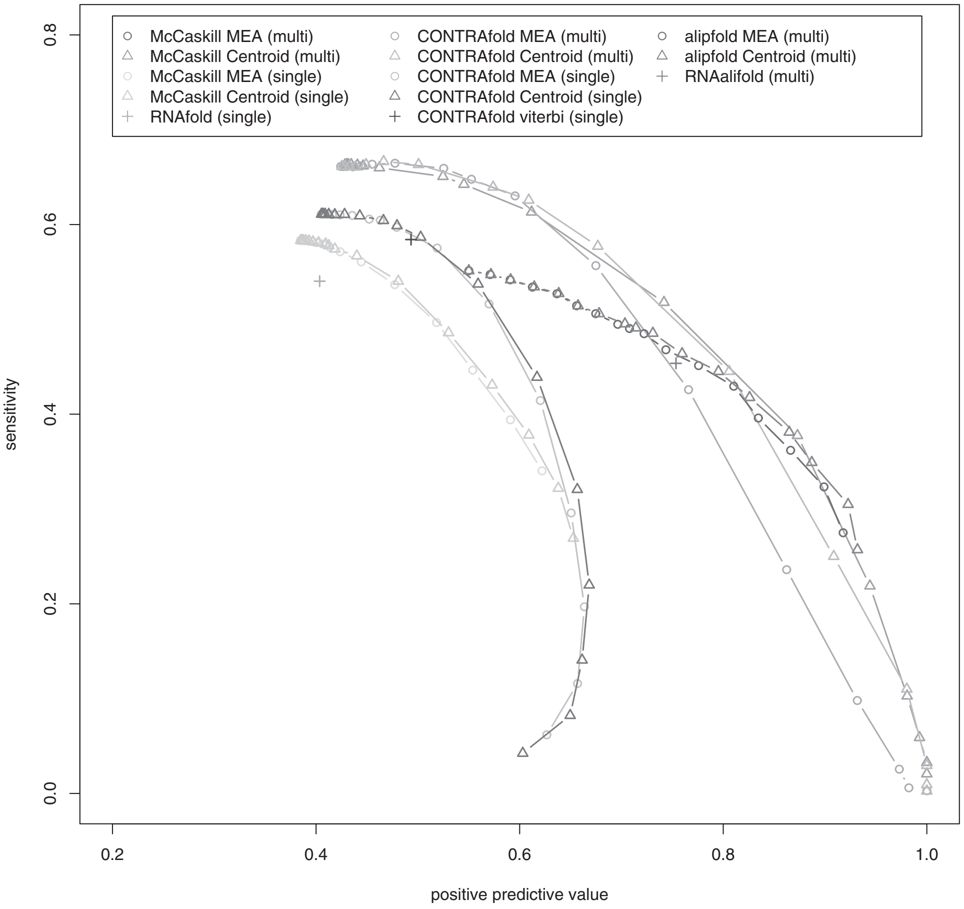


Fig. 6. Performance of common secondary structure prediction with ProbCons alignments. See the main text for a description of each estimator. We have plotted the curves at $\gamma \in \{2^k : -5 \leq k \leq 10, k \in \mathbb{Z}\} \cup \{6\}$. Also see Supplementary Figure S3, S4, S5 and S6 for the results corresponding to a reference alignment or an alignment produced by other aligners [MXSCARNA (Tabei *et al.*, 2008), MAFFT (Katoh *et al.*, 2005) and ClustalW (Thompson *et al.*, 1994)].

terms in (C.4), which is provided in the Supplementary Matrial, in such way that less accurate alignments emphasize the difference between the McCaskill-MEA estimator and the averaged γ-centroid estimator. We also point out that the difference in performance between the averaged γ-centroid estimator and the McCaskill-MEA estimator for common secondary structure prediction is larger than that between the γ-centroid estimator and the MEA estimator for secondary structure prediction [i.e. $|(a)-(b)| < |(i)-(iv)|$[4] and $|(c)-(d)| < |(ii)-(v)|$). It is also interesting that the difference between the McCaskill model and the CONTRAfold model is smaller for common secondary structure prediction than for secondary structure prediction. We consider this to be attributable to the fact that the peaks around 1 in the histograms of the base paring probabilities of the true base pairs and the false base pairs with the McCaskill model (the top two histograms in Supplementary Figure S13) disappear in the histograms of the averaged base-pairing probabilities in the McCaskill model (top two histograms in Supplementary Figure S12). As we stated in the previous section, we think that the peak in the histogram of the *false* base-pairing probabilities of the McCaskill model prevents the γ-centroid estimator and the MEA estimator for the secondary structure prediction from increasing PPV and sensitivity because potential base-pairs whose probability is close to 1 can be easily predicted as base pairs. We consider that the high probabilities of false base pairs are averaged out by averaging base-pairing probability matrix because the McCaskill model randomly gives high probability to false base pairs, and this leads to make the difference of the performance between the CONTRAfold model and the McCaskill model smaller on the common secondary structure prediction.

From Table 2, we see that the values of γ of the averaged γ-centroid estimator which provide the highest MCC value are $\gamma=2$ for the McCaskill model and $\gamma=4$ for the CONTRAfold model, respectively. This table also shows that the value of the γ parameter for the proposed method is important for its performance with respect to common secondary structure prediction. We also conducted the experiments for alignments which contain 2, 4, 6, 8 or 10 sequences (the dataset is created by randomly selecting sequences from the main dataset). Figure 7 shows that the fewer sequences there are in the alignment, the better the averaged γ-centroid is, compared with the MEA estimator.

[4] This notation means that the difference in performance between (i) and (iv) is larger than that between (a) and (b).

Table 2. MCC of the averaged γ-centroid estimators for the common secondary structure prediction

	McCaskill model					CONTRAfold model				
γ	Ref	Clustalw	Mafft-ginsi	Mxscarna	Probcons	Ref	Clustalw	Mafft-ginsi	Mxscarna	Probcons
0.03125	0.153	0.137	0.128	0.156	0.143	0.000	0.000	0.000	0.000	0.000
0.0625	0.199	0.169	0.171	0.205	0.181	0.060	0.050	0.050	0.050	0.050
0.125	0.275	0.217	0.221	0.276	0.242	0.103	0.096	0.096	0.103	0.096
0.25	0.380	0.276	0.288	0.369	0.317	0.209	0.165	0.163	0.207	0.172
0.5	0.563	0.391	0.403	0.524	0.454	0.394	0.279	0.296	0.373	0.328
1	0.707	0.493	0.503	0.630	0.573	0.600	0.405	0.430	0.537	0.476
2	**0.790**	**0.535**	0.543	**0.671**	**0.619**	0.744	0.515	0.537	0.644	0.598
4	0.783	0.534	**0.546**	0.659	0.611	**0.795**	**0.545**	**0.578**	**0.678**	**0.624**
6	0.757	0.512	0.529	0.646	0.591	0.785	0.531	0.569	0.669	0.616
8	0.748	0.507	0.524	0.64	0.583	0.777	0.526	0.561	0.662	0.605
16	0.708	0.476	0.497	0.617	0.551	0.734	0.493	0.535	0.639	0.575
32	0.696	0.467	0.488	0.607	0.542	0.710	0.478	0.518	0.622	0.556
64	0.691	0.464	0.485	0.603	0.539	0.696	0.467	0.508	0.609	0.544
128	0.686	0.462	0.481	0.600	0.535	0.688	0.460	0.501	0.603	0.537
256	0.684	0.461	0.480	0.598	0.534	0.683	0.457	0.498	0.599	0.533
512	0.682	0.460	0.478	0.598	0.533	0.68	0.456	0.496	0.597	0.531
1024	0.681	0.459	0.477	0.596	0.531	0.678	0.454	0.494	0.595	0.529

We use Kiryu *et al.*'s (2007) dataset.

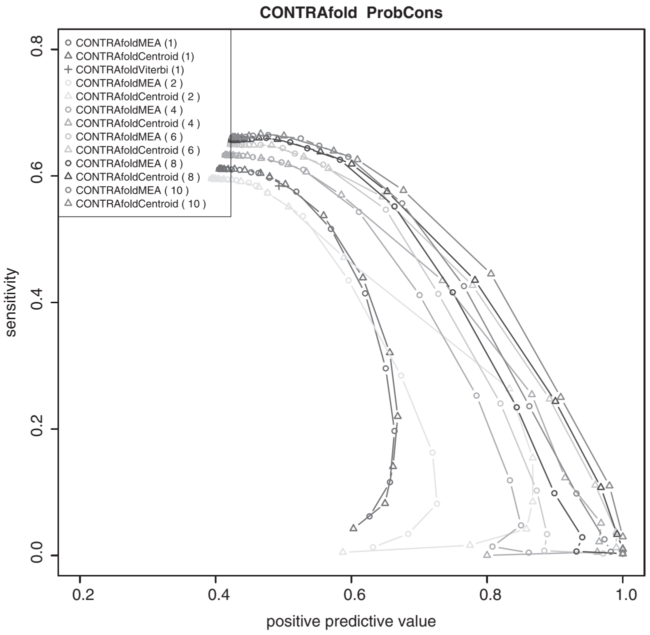


Fig. 7. Performance of common secondary structure prediction of the averaged γ-centroid estimators and the MEA estimators for alignments containing 2, 4, 6, 8 or 10 sequences with ProbCons alignments and the CONTRAfold model. Other combinations of alignment tools and probability distributions are shown in the Supplementary Data.

4 DISCUSSION AND CONCLUSION

In this article, we have designed novel estimators, namely the γ-centroid estimator for secondary structure prediction from individual RNA sequences and the averaged γ-centroid estimator for common secondary structure prediction from multiple alignments of RNA sequences. Our principles for designing an estimator are as follows: (i) The estimators should be suitable evaluation measures, but (ii) they must be able to be computed efficiently. For example, we can design estimators which maximize the expectation of MCC, but they are not computationally efficient. We believe that the choices of the gain function presented in this article are the best from this point of view. Our estimators are extensions of the centroid estimators proposed in Ding *et al.* (2005) and Carvalho and Lawrence (2008) with a parameter γ, which controls the balance between the sensitivity and PPV as in the case of the MEA-based estimator (Do *et al.*, 2006; Kiryu *et al.*, 2007). It has been shown that the proposed estimators reflect the accuracy measures more precisely than MEA-based estimators presented in previous studies, the MEA estimators (Do *et al.*, 2006) for secondary structure prediction and the McCaskill-MEA (Kiryu *et al.*, 2007) estimator for common secondary structure prediction since those estimators include unnecessary terms which never contribute to the evaluation of the accuracy. It has also been shown in computational experiments that the proposed estimators outperform MEA-based estimators when the same scoring models of the secondary structures are used (the scoring models are interpreted as either the probability distributions over the structures or the energy models of the structures). The combinations of estimators and scoring models which had the highest performance were the γ-centroid estimators with the CONTRAfold model for secondary structure prediction and the averaged γ-centroid estimators with the CONTRAfold model or the McCaskill model for common secondary structure prediction.

The proposed estimators are applicable to a wide variety of estimation problems on binary spaces in bioinformatics, as in the case of centroid estimators (Carvalho and Lawrence, 2008). This provides a generalized framework for the design of highly accurate estimators for various problems, where the balance between the sensitivity and PPV is controlled by γ. For example, the estimators used in the ProbCons program (Do *et al.*, 2005) are special cases of the averaged γ-centroid estimator.

ACKNOWLEDGEMENTS

The authors thank L. E. Carvalho and C. E. Lawrence for valuable comments. The authors are also grateful to T. Kato for reading the manuscript and providing useful comments, as well as to our colleagues at the Computational Biology Research Center (CBRC) for fruitful discussions. Also, the constructive comments of the anonymous reviewers are greatly appreciated.

Funding: New Energy and Industrial Technology Development Organization (NEDO) of Japan (to 'Functional RNA project); 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**, 19–28.

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

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, UK.

Griffiths-Jones,S. *et al*. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.

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.

Katoh,K. *et al*. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

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

Kiryu,H. *et al*. (2008) Rfold: an exact algorithm for computing local base pairing probabilities. *Bioinformatics*, **24**, 367–373.

Mattick,J. (2005) The functional genomics of noncoding RNA. *Science*, **309**, 1527–1528.

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

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

Prabhakar,S. *et al*. (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science*, **314**, 786.

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

Thompson,J.D. *et al*. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Venkatesh,B. *et al*. (2006) Ancient noncoding elements conserved in the human genome. *Science*, **314**, 1892.

Washietl,S. *et al*. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.

Zaratiegui,M. *et al*. (2007) Noncoding RNAs and gene silencing. *Cell*, **128**, 763–776.

Zuker,M. and Stiegler,P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.