Supplementary Information:
Improving the accuracy of predicting secondary structure for aligned RNA sequences

Michiaki Hamada, Kengo Sato and Kiyoshi Asai

A Detailed descriptions using mathematics

A.1 Notations used in this paper

The following notations are used in this paper. For an RNA sequence $x$, $x_i$ is the $i$-th base in $x$, and $|x|$ is the length of $x$. A secondary structure $\theta$ of an RNA sequence $x$ is represented as an upper triangular binary matrix $\theta = \{\theta_{ij}\}_{i,j}$, where $\theta_{ij}$ is 1 if $x_i$ and $x_j$ form a base-pair and 0 otherwise. We denote by $S(x)$ the space of all secondary structures of an RNA sequence $x$. For a multiple alignment $A$ of RNA sequences, $|A|$ is the length of the alignment $A$. Since $S(x)$ only depends on the length of the RNA sequence $x$, we denote by $S(A)$ the space of secondary structures of an RNA sequence whose length is equal to $|A|$.

There are several probability distributions of secondary structures of a given RNA sequence: the models based on machine learning used in CONTRAfold [3] and Simfold [1], or the energy-based models used in Mfold [16], RNAfold [5] and RNAstructure [11]. In this paper, we denote the CONTRAfold and McCaskill models on $S(x)$ by $p^{(contra)}(\theta|x)$ and $p^{(mcc)}(\theta|x)$, respectively. Moreover, there are probability distributions for the common secondary structures of a given alignment $A$: (i) the RNAalipfold model (denoted by $p^{(alipfold)}(\theta|A)$) is a probabilistic version of RNAalifold, which is based on the free energy of each secondary structure in the alignment and the covariance bonus, and (ii) the Pfold model [10, 9] (denoted by $p^{(pfold)}(\theta|A)$), which is based on stochastic context free grammars (SCFGs) and phylogenetic information.

A.2 Problem: secondary structure prediction for aligned RNA sequences

The problem considered in this paper, which is usually called common (consensus or representative) secondary structure prediction, can be stated as follows.

**Problem 1** Given a multiple alignment $A$ of RNA sequences, then predict a secondary structure $y$ in $S(A)$.

A.3 Two Evaluation Processes for Problem 1

A.3.1 Evaluation Process 1 (Fig. S1)

The predicted common secondary structure $y$ in Problem 1 is compared with the reference common secondary structure $\theta$. Then three accuracy measures, sensitivity (SEN), positive predictive value (PPV) and Matthews Correlation Coefficient (MCC) are used: $\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}$, $\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$ and

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$
Figure S1: Evaluation Process 1. We assume that the reference (common) secondary structure of the input multiple alignment is given. The comparison between a predicted structure and the reference structure is based on the base-pairs in the secondary structures (using the sensitivity (SEN), the positive predictive value (PPV) and the Matthews Correlation Coefficient (MCC) with respect to base-pairs).

Figure S2: Evaluation Process 2. We assume that the reference secondary structure of every RNA sequence in the input alignment is given. The comparison between every mapped structure and the reference structure is based on the base-pairs in the secondary structures using SEN, PPV, MCC.

where TP, TN, FP, and FN are the numbers of true positive base-pairs, true negative base-pairs, false positive base-pairs and false negative base-pairs, respectively. These are calculated from

\[
TP = TP(\theta, y) = \sum_{i<j} I(y_{ij} = 1)I(\theta_{ij} = 1),
\]

\[
TN = TN(\theta, y) = \sum_{i<j} I(y_{ij} = 0)I(\theta_{ij} = 0),
\]

\[
FP = FP(\theta, y) = \sum_{i<j} I(y_{ij} = 1)I(\theta_{ij} = 0)
\]

and

\[
FN = FN(\theta, y) = \sum_{i<j} I(y_{ij} = 0)I(\theta_{ij} = 1)
\]  

where \(I(\cdot)\) is the indicator function that takes a value of 1 or 0 depending on whether the condition constituting its argument is true or false.

A.3.2 Evaluation Process 2 (Fig. S2)

The predicted common secondary structure \(y\) is evaluated as follows. First, \(y\) is mapped onto each RNA sequence in \(A\). Then, 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. Second, TP, TN, FP, and FN are calculated for each mapped secondary structure \(y^{(map)}\) with respect to the reference secondary structure \(\theta\) using the formulas in Eq. (S1). Finally, SEN, PPV and MCC are calculated for the sum of TP, TN, FP, and FN over all the RNA sequences in \(A\).

A.4 General form of MEA-based estimator for Problem 1

A.4.1 MEA-based estimator (E1) that is consistent with Evaluation Process 1

In order to introduce the estimator that is consistent with Evaluation Process 1, we assume that there is a probability distribution \(p(\theta|A)\) of the common secondary structures of \(A\) (i.e., \(S(A)\)) and a gain function \(G(\theta, y)\) between two secondary structures \(\theta\) (the reference structure) and \(y\) (predicted structure) whose
lengths are equal to the length of the alignment $A$ (Fig. 1 in the main manuscript). Then, the estimator that maximizes the expected gain is

$$\hat{y} = \arg \max_{y \in \mathcal{S}(A)} \sum_{\theta \in \mathcal{S}(A)} G(\theta, y)p(\theta|A). \quad (S2)$$

A straight-forward way for obtaining $p(\theta|A)$ is to define $p(\theta|A) = p^{(\text{fold})}(\theta|A)$ or $p^{(\text{alifold})}(\theta|A)$. On the other hand, the gain function $G(\theta, y)$ should be consistent with the accuracy measures: SEN, PPV and MCC. Three concrete examples of $G(\theta, y)$ are shown in later sections.

### A.4.2 MEA-based estimator (E2) that is consistent with Evaluation Process 2

As shown in Fig. 2 in the main manuscript, we assume that we obtain a set of probability distributions $\{p_x(\theta|A)\}_{x \in A}$ where $p_x(\theta|A)$ is a probability distribution, given the alignment $A$, of the secondary structures of the RNA sequence $x \in A$. (Note that $x$ might contain several gaps.) We consider a gain function $G(\theta, y)$ between two secondary structures, $\theta$ (reference) and $y$ (prediction), having the same length. These assumptions are consistent with Evaluation Process 2.

Then, we obtain the following general estimator for common secondary structure.

$$\hat{y} = \arg \max_{y \in \mathcal{S}} \sum_{x \in A} \left[ \sum_{\theta \in \mathcal{S}} G(\theta, y)p_x(\theta|A) \right] \quad (S3)$$

where $\mathcal{S} = \mathcal{S}(A)$ (note that $\mathcal{S} = \mathcal{S}(A) = \mathcal{S}(x)$ for all $x \in A$). In other words, the estimator of Eq. (S3) predicts the common secondary structure by maximizing the sum of the expected gain over all the RNA sequences in the alignment $A$. A straight-forward way for obtaining $p_x(\theta|A)$ is to define $p_x(\theta|A) = p^{(\text{mec})}(\theta|x)$ or $p_x(\theta|A) = p^{(\text{contra})}(\theta|x)$.

### A.4.3 Relation between the estimator (E1) and (E2)

The two estimators (E1) in Eq. (S2) and (E2) in Eq. (S3) are closely related to each other. It is easily seen that the estimator (E2) of Eq. (S3) is equivalent to

$$\hat{y} = \arg \max_{y \in \mathcal{S}^c} \sum_{\theta \in \mathcal{S}^c} G(\theta, y)\overline{p}(\theta|A) \quad (S4)$$

where $\overline{p}(\theta|A)$ is the averaged probability distribution

$$\overline{p}(\theta|A) := \frac{1}{n} \sum_{x \in A} p_x(\theta|A). \quad (S5)$$

where $n$ is the number of sequences in $A$. Therefore, the estimator (E2) of Eq. (S3) is equivalent to the estimator (E1) of Eq. (S2) if we take $\overline{p}(\theta|A)$ as the probability distribution of the common secondary structures. Conversely, the estimator (E1) of Eq. (S2) is equivalent to the estimator (E2) of Eq. (S3) if we take $p_x(\theta|A)$ in Eq. (S3) to be identical to the distribution $p(\theta|A)$ in Eq. (S2) for all $x \in A$. This implies that the estimator (E1) of Eq. (S2) is unsuitable for Evaluation Process 2 because it is not necessary that all $p_x(\theta|A)$ are identical to the distribution in Eq. (S3).

### A.5 State-of-the-art algorithms with MEA-based estimators

In this section, we describe several state-of-the-art algorithms by using the context of MEA-based estimators. This representation is useful in our classification of the algorithms (Table 1 in the main manuscript).

#### A.5.1 CentroidAlifold [4]

CentroidAlifold is considered as the estimator (E2) of Eq. (S3) as follows. For $\gamma > 0$, the gain function $G(\theta, y)$ in Eq. (S3) is given by

$$G^{(\text{centroid})}_\gamma(\theta, y) = \gamma \cdot TP(\theta, y) + TN(\theta, y) \quad (S6)$$

where $TP(\theta, y)$ and $TN(\theta, y)$ are defined in Eq. (S1). As shown in [4], this gain function is consistent with the widely used accuracy measures of secondary structure prediction, namely, SEN, PPV and MCC with
A.5.2 RNAalifold [2, 6]

RNAalifold [2, 6] maximizes a score that combines the free energy of each RNA sequence in the input alignment with the covariance bonus. Also, by using the score, a probability distribution of common secondary structures is introduced (we call it RNAalipfold model in the manuscript).

To compare RNAalifold with the other MEA-based algorithms, we give the following (in-direct) implications. From a probabilistic viewpoint, RNAalifold is equivalent to the maximum likelihood (ML) estimator for the probability distribution of the (common) secondary structures given by the RNAalipfold model:

\[ \hat{y} = \arg \max_{y \in S(A)} p^{(alipfold)}(y|A). \]

This is naturally considered as the estimator (E1) of Eq. (S2) as follows. The gain function \( G \) in Eq. (S2) is given by

\[ G^{(\delta)}(\theta, y) = \delta(\theta, y) \]  

where \( \delta \) is the delta function that gives 1 only when the secondary structure \( \theta \) is exactly the same as \( y \), and the probability distribution \( p(\theta|A) \) in Eq. (S2) is given by \( p^{(alipfold)}(\theta|A) \).

A.5.3 RNAalipfold-Centroid [2, 4]

We can employ the MEA-based method (or posterior decoding method), such as \( \gamma \)-centroid estimator, with respect to RNAalipfold model.

RNAalipfold-Centroid is the \( \gamma \)-centroid estimator [4] with RNAalipfold model. This is equivalent to the estimator (E1) of Eq. (S2) in which \( G(\theta, y) = G^{(centroid)}(\gamma, \theta, y) \) (Eq. (S6)) and \( p(\theta|A) = p^{(alipfold)}(\theta|A) \) (RNAalipfold model).

A.5.4 PETfold [13]

Although the authors did not state it explicitly, the main part of PETfold can be considered as the estimator (E2) in Eq. (S3) as follows. First, the probability distribution \( p_x(\theta|A) \) in Eq. (S3) is defined by a mixture of the McCaskill and Pfold models:

\[ p_x(\theta|A) = \frac{\beta}{1+\beta} p^{(mcc)}(\theta|x) + \frac{1}{1+\beta} p^{(pfold)}(\theta|A) \]

where \( \beta \) is the same parameter that appears in the original paper [13]. (Precisely speaking, \( p^{(mcc)}(\theta|x) \) is extended to the sequence with gaps by using Eq. (S7).) As a default, equal weight to \( p^{(pfold)} \) and \( p^{(mcc)} \) is given (\( \beta = 1 \)). Second, the gain function \( G(\theta, y) \) in Eq. (S3) is given by

\[ G^{(contra)}_i(\theta, y) = \sum_{i=1}^{n} \gamma \sum_{j \neq i} I(\theta^*_{ij} = 1) I(y^*_{ij} = 1) + \prod_{j:j \neq i} I(\theta^*_{ij} = 0) I(y^*_{ij} = 0) \]  

where \( \theta^* \) and \( y^* \) are the symmetric extensions of (the upper triangular matrices) \( \theta \) and \( y \), respectively. (i.e., \( \theta^*_{ij} = \theta_{ji} \) for \( i < j \) and \( \theta^*_{ij} = \theta_{ji} \) for \( j < i \). The extension for \( y \) is similar.) This gain function is the same as the one used in CONTRAfold [3] for conventional secondary structure prediction. Then, the estimator (E2) of Eq. (S3) is equivalent to the main part of the PETfold algorithm. In addition, PETfold employs some heuristics with the above estimator that select reliably conserved substructures before estimating the common secondary structure [13].
A.5.5 Pfold [10, 9]

The current version of Pfold [10] is equivalent to the estimator (E1) of Eq. (S2), in which $G(\theta, y) = G_{1}^{(\text{contra})}(\theta, y)$ (i.e. $G_{\gamma}^{(\text{contra})}(\theta, y)$ with $\gamma = 1$) and $p(\theta|A) = p^{(\text{pfold})}(\theta|A)$ (Pfold model). (Note that the previous version of Pfold [9] employs $G^{\delta}$ as the gain function.)

A.5.6 Pfold-Centroid [4, 10]

Pfold-Centroid is the $\gamma$-centroid estimator with Pfold model, that is, the estimator (E1) of Eq. (S2) where the gain function is given by $G_{\gamma}^{(\text{centroid})}(\theta, y)$ and the probability distribution is given by $p^{(\text{pfold})}(\theta|A)$.

A.5.7 McCaskill-MEA [8]

McCaskill-MEA can be naturally considered as the estimator (E2) of Eq. (S3) where the gain function $G(\theta, y)$ is given by $G_{\gamma}^{(\text{contra})}(\theta, y)$ (Eq. (S10)) and the probability distribution $p_x(\theta|A)$ is given by Eq. (S7) with the McCaskill model [12]. Note that Hamada et al have shown that the McCaskill-MEA is consistently worse than CentroidAlifold in computational experiments [4] although the difference is only the gain functions.

B Supplementary Figures and Table
Figure S3: Performance of the representative secondary structure prediction with the alignments produced by ClustalW [15]
Figure S4: Performance of the representative secondary structure prediction with the alignments produced by MAFFT [7]
Figure S5: Performance of the representative secondary structure prediction with the alignments produced by MXSCARNA [14]
Figure S6: The performances of CentroidAlifold with various values of the weight parameter (i.e. $w = 0, 0.1, 0.2, \ldots, 0.9, 1$ in Eq. (2) in the main manuscript). We used alignments produced by ClustalW.
Figure S7: The performances of CentroidAlifold with various values of the weight parameter (i.e. $w = 0, 0.1, 0.2, \ldots, 0.9, 1$ in Eq. (2) in the main manuscript). We used alignments produced by ProbconsRNA.
Figure S8: The performances of CentroidAlifold with various values of the weight parameter (i.e. $w = 0, 0.1, 0.2, \ldots, 0.9, 1$ in Eq. (2) in the main manuscript). We used alignments produced by MAFFT.
Figure S9: The performances of CentroidAlifold with various values of the weight parameter (i.e. $w = 0, 0.1, 0.2, \ldots, 0.9, 1$ in Eq. (2) in the main manuscript). We used alignments produced by MXSCARNA.
Figure S10: The performances of CentroidAlifold with various values of the weight parameter (i.e. $w = 0, 0.1, 0.2, \ldots, 0.9, 1$ in Eq. (2) in the main manuscript). We used the reference alignments.
| $\gamma$ | clustalw | mxscarna | probcons | mafft-qinsi | ref |
|-------|---------|---------|---------|-----------|-----|
|       | SEN    | PPV    | MCC     | SEN    | PPV    | MCC     | SEN    | PPV    | MCC     | SEN    | PPV    | MCC     | SEN    | PPV    | MCC     |
| 0.03125 | 0.027 | 0.997 | 0.163 | 0.041 | 0.979 | 0.199 | 0.031 | 1.000 | 0.176 | 0.035 | 0.982 | 0.185 | 0.038 | 1.000 | 0.195 |
| 0.0625 | 0.045 | 0.996 | 0.211 | 0.079 | 0.957 | 0.275 | 0.057 | 0.993 | 0.237 | 0.065 | 0.982 | 0.252 | 0.073 | 1.000 | 0.270 |
| 0.125  | 0.079 | 0.981 | 0.278 | 0.165 | 0.899 | 0.384 | 0.107 | 0.982 | 0.324 | 0.128 | 0.953 | 0.349 | 0.161 | 0.994 | 0.399 |
| 0.25   | 0.169 | 0.932 | 0.396 | 0.351 | 0.857 | 0.548 | 0.242 | 0.939 | 0.476 | 0.270 | 0.898 | 0.492 | 0.372 | 0.973 | 0.601 |
| 0.5    | 0.287 | 0.861 | 0.497 | 0.539 | 0.810 | 0.660 | 0.408 | 0.885 | 0.600 | 0.452 | 0.845 | 0.617 | 0.614 | 0.936 | 0.774 |
| 1.0    | 0.365 | 0.813 | 0.544 | 0.612 | 0.778 | 0.689 | 0.490 | 0.848 | 0.644 | 0.547 | 0.795 | 0.659 | 0.751 | 0.917 | 0.829 |
| 2.0    | 0.397 | 0.735 | 0.540 | 0.646 | 0.733 | 0.687 | 0.535 | 0.792 | 0.651 | 0.592 | 0.752 | 0.667 | 0.791 | 0.877 | 0.832 |
| 4.0    | 0.451 | 0.657 | 0.543 | 0.677 | 0.684 | 0.680 | 0.579 | 0.730 | 0.649 | 0.632 | 0.701 | 0.665 | 0.822 | 0.835 | 0.828 |
| 6.0    | 0.501 | 0.595 | 0.545 | 0.704 | 0.642 | 0.671 | 0.624 | 0.664 | 0.643 | 0.667 | 0.636 | 0.650 | 0.844 | 0.787 | 0.814 |
| 8.0    | 0.526 | 0.561 | 0.542 | 0.717 | 0.618 | 0.665 | 0.643 | 0.629 | 0.635 | 0.681 | 0.604 | 0.640 | 0.856 | 0.760 | 0.806 |
| 16.0   | 0.566 | 0.470 | 0.515 | 0.739 | 0.566 | 0.646 | 0.675 | 0.546 | 0.606 | 0.704 | 0.539 | 0.615 | 0.881 | 0.680 | 0.774 |
| 32.0   | 0.578 | 0.412 | 0.487 | 0.748 | 0.523 | 0.624 | 0.686 | 0.487 | 0.577 | 0.716 | 0.494 | 0.594 | 0.893 | 0.610 | 0.737 |
| 64.0   | 0.580 | 0.396 | 0.478 | 0.750 | 0.511 | 0.618 | 0.689 | 0.472 | 0.569 | 0.716 | 0.475 | 0.582 | 0.893 | 0.592 | 0.726 |
| 128.0  | 0.581 | 0.388 | 0.473 | 0.751 | 0.503 | 0.613 | 0.690 | 0.463 | 0.564 | 0.716 | 0.465 | 0.576 | 0.894 | 0.581 | 0.720 |
| 256.0  | 0.582 | 0.384 | 0.471 | 0.751 | 0.497 | 0.610 | 0.690 | 0.457 | 0.560 | 0.716 | 0.459 | 0.572 | 0.894 | 0.575 | 0.716 |
| 512.0  | 0.583 | 0.380 | 0.469 | 0.751 | 0.494 | 0.608 | 0.691 | 0.454 | 0.559 | 0.716 | 0.456 | 0.570 | 0.895 | 0.571 | 0.714 |
| 1024.0 | 0.582 | 0.379 | 0.468 | 0.750 | 0.492 | 0.606 | 0.691 | 0.452 | 0.557 | 0.716 | 0.454 | 0.568 | 0.895 | 0.567 | 0.711 |
References

[1] M. Andronescu, A. Condon, H. Hoos, D. Mathews, and K. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23:19–28, Jul 2007.

[2] S. Bernhart, I. Hofacker, S. Will, A. Gruber, and P. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, Nov 2008.

[3] C. Do, D. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22:e90–98, Jul 2006.

[4] M. Hamada, H. Kiryu, K. Sato, T. Mituyama, and K. Asai. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, 25:465–473, Feb 2009.

[5] I. Hofacker, W. Fontana, P. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[6] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–1066, Jun 2002.

[7] K. Katoh, K. Kuma, H. Toh, and T. Miyata. Maft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–518, 2005.

[8] H. Kiryu, T. Kin, and K. Asai. Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics*, 23:434–441, Feb 2007.

[9] B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15:446–454, Jun 1999.

[10] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, Jul 2003.

[11] D. Mathews, M. Disney, J. Childs, S. Schroeder, M. Zuker, and D. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, 101:7287–7292, May 2004.

[12] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, May 1990.

[13] S. Seemann, J. Gorodkin, and R. Backofen. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, 36:6355–6362, Nov 2008.

[14] Y. Tabei, H. Kiryu, T. Kin, and K. Asai. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, 9:33, Jan 2008.

[15] J. D. Thompson, D. G. Higgins, and T. J. Gibson. 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(22):4673–4680, Nov 1994.

[16] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–3415, Jul 2003.