

Generalized Centroid Estimators in Bioinformatics*

Michiaki Hamada^{1,2†}, Hisanori Kiryu¹, Wataru Iwasaki¹, Kiyoshi Asai^{1,2}

¹the University of Tokyo, ²CBRC/AIST

Abstract

In a number of estimation problems in bioinformatics, accuracy measures of the target problem are usually given, and it is important to design estimators that are suitable to those accuracy measures. However, there is often a discrepancy between an employed estimator and a given accuracy measure of the problem. In this study, we introduce a general class of efficient estimators for estimation problems on high-dimensional binary spaces, which represent many fundamental problems in bioinformatics. Theoretical analysis reveals that the proposed estimators generally fit with commonly-used accuracy measures (e.g. sensitivity, PPV, MCC and F-score) as well as it can be computed efficiently in many cases, and cover a wide range of problems in bioinformatics from the viewpoint of the principle of maximum expected accuracy (MEA). It is also shown that some important algorithms in bioinformatics can be interpreted in a unified manner. Not only the concept presented in this paper gives a useful framework to design MEA-based estimators but also it is highly extendable and sheds new light on many problems in bioinformatics.

Contents

1	Introduction	2
2	Materials and Methods	3
3	Results	5
3.1	γ -centroid estimator: generalized centroid estimator	5
3.2	Generalized centroid estimators for representative prediction	8
3.3	Estimators based on marginal probabilities	9
4	Discussion	11
4.1	Properties of the γ -centroid estimator	11
4.2	How to determine the parameter in γ -centroid estimator	12
4.3	Accuracy measures and computational efficiency	12
4.4	Probability distributions are not always defined on predictive space	13
4.5	Application of γ -centroid estimator to cluster centroid	13
4.6	Conclusion	14

*This is a corrected version of the published paper: *PLoS ONE* 6(2):e16450, 2011. The original version is available from <http://www.plosone.org/article/info:doi/10.1371/journal.pone.0016450>. Note that there are several typos in the original version which is not in the accepted manuscript (we had no chance for proof reading of the published paper).

[†]To whom correspondence should be addressed. Tel.: +81-3-5281-5271; Fax: +81-3-5281-5331; E-mail: mhamada@k.u-tokyo.ac.jp

A	Appendices	16
A.1	Discrete (binary) spaces in bioinformatics	16
A.1.1	The space of alignments of two biological sequences: $\mathcal{A}(x, x')$	16
A.1.2	The space of secondary structures of RNA: $\mathcal{S}(x)$	16
A.1.3	The space of phylogenetic trees: $\mathcal{T}(S)$	17
A.2	Probability distributions on discrete spaces	17
A.2.1	Probability distributions $p^{(a)}(\theta x, x')$ on $\mathcal{A}(x, x')$	17
A.2.2	Probability distributions $p^{(s)}(\theta x)$ on $\mathcal{S}(x)$	18
A.2.3	Probability distributions $p^{(t)}(\theta S)$ on $\mathcal{T}(S)$	18
A.3	Evaluation measures defined using TP, TN, FP and FN	18
A.4	Schematic diagrams of representative and approximated γ -type estimators	19
A.5	Applications in bioinformatics	20
A.5.1	Pairwise alignment of biological sequences (Problem 1)	20
A.5.2	Secondary structure prediction of an RNA sequence (Problem 2)	22
A.5.3	Estimation of phylogenetic trees (Problem 4)	23
A.5.4	Alignment between two <i>alignments</i> of biological sequences	24
A.5.5	Common secondary structure prediction from a multiple alignment of RNA sequences	25
A.5.6	Pairwise alignment using homologous sequences	26
A.5.7	RNA secondary structure prediction using homologous sequences	28
A.5.8	Pairwise alignment of <i>structured</i> RNAs	29
A.6	Proofs	31
A.6.1	Proof of Theorem 1	31
A.6.2	Proof of Theorem 2	31
A.6.3	Proof of Theorem 3	32
A.6.4	Proof of Corollary 1	32
A.6.5	Proof of Proposition 1	32
A.6.6	Derivation of Eq. (14)	33

1 Introduction

In estimation problems in bioinformatics, the space of solutions is generally large and often high-dimensional. Among them, a number of fundamental problems in bioinformatics, such as alignment of biological sequences, prediction of secondary structures of RNA sequences, prediction of biological networks, and estimation of phylogenetic trees, are classified into estimation problems whose solutions are in a high-dimensional binary space. Such problems are generally difficult to solve, and the estimates are often unreliable.

The popular solutions for these problems, such as for the secondary structure of RNA with minimum free energy, are the maximum likelihood (ML) estimators. The ML estimator maximizes the probability that the estimator is exactly correct, but that probability is generally very small. Noticing the drawbacks of the ML estimators, Carvalho and Lawrence have proposed the *centroid estimator*, which represents an ensemble of all the possible solutions and minimizes the expected Hamming loss of the prediction [1].

In this paper, we conduct a theoretical analysis of estimation problems in high-dimensional binary space, and present examples and solutions in bioinformatics. The theories in this paper provide a unified framework for designing superior estimators for estimation problems in bioinformatics. The estimators discussed in this paper, including the ML estimator and the centroid estimator, are formalized as maximum expected gain (MEG) estimators, which maximize the estimator-specific gain functions with respect to the given probability distribution. The objective of the estimation is not always to find the exact solution with an extremely small probability or to find the solution with the minimum Hamming loss, but rather to find the most accurate estimator. Therefore, we adopt the principle of maximum expected accuracy (MEA),

which has been successfully applied to various problems in bioinformatics, such as the alignment of biological sequences [2–4], the secondary structure prediction of RNA [5–8] and other applications [9–11].

Theoretical analysis, however, shows that those MEA estimators are not always robust with respect to accuracy measures. To address this, we previously proposed the γ -centroid estimator in a few specific problems [4, 12]. In this paper, in order to make the γ -centroid estimator easily applicable to other estimation problems, we introduce an abstract form of the γ -centroid estimator, which is defined on general binary spaces and designed to fit to the commonly used accuracy measures. The γ -centroid estimator is a generalization of the centroid estimator, and offers a more robust framework for estimators than the previous estimators. We extend the theory of maximum expected gain (MEG) estimators and γ -centroid estimators for two advanced problems: the estimators that represent the common solutions for multiple entries, and the estimators for marginalized probability distributions.

2 Materials and Methods

Problem 1 (Pairwise alignment of two biological sequences) *Given a pair of biological (DNA, RNA, protein) sequences x and x' , predict their alignment as a point in $\mathcal{A}(x, x')$, the space of all the possible alignments of x and x' .*

Problem 2 (Prediction of secondary structures of RNA sequences) *Given an RNA sequence x , predict its secondary structure as a point in $\mathcal{S}(x)$, the space of all the possible secondary structures of x .*

A point in $\mathcal{A}(x, x')$, can be represented as a binary vector of $|x||x'|$ dimensions by denoting the aligned bases across the two sequences as "1" and the remaining pairs of bases as "0". A point in $\mathcal{S}(x)$ can also be represented as a binary vector of $|x|(|x| - 1)/2$ dimensions, which represent all the pairs of the base positions in x , by denoting the base pairs in the secondary structures as "1". In each problem, the predictive space ($\mathcal{A}(x, x')$ or $\mathcal{S}(x)$) is a subset of a binary space ($\{0, 1\}^{|x||x'|}$ or $\{0, 1\}^{|x|(|x|-1)/2}$) because the combinations of aligned bases or base pairs are restricted (see "Discrete (binary) spaces in bioinformatics" (Section A.1) in Appendices for more formal definitions). Therefore, Problem 1 and Problem 2 are special cases of the following more general problem:

Problem 3 (Estimation problem on a binary space) *Given a data set D and a predictive space Y (a set of all candidates of a prediction), which is a subset of n -dimensional binary vectors $\{0, 1\}^n$, that is, $Y \subset \{0, 1\}^n$, predict a point y in the predictive space Y .*

Not only Problem 1 and Problem 2 but also a number of other problems in bioinformatics are formulated as Problem 3, including the prediction of biological networks and the estimation of phylogenetic trees (Problem 4).

To discuss the stochastic character of the estimators, the following assumption is introduced.

Assumption 1 (Existence of probability distribution) *In Problem 3, there exists a probability distribution $p(y|D)$ on the predictive space Y .*

For Problem 3 with Assumption 1, we have the following Bayesian maximum likelihood (ML) estimator.

Definition 1 (Bayesian ML estimator [1]) *For Problem 3 with Assumption 1, the estimator*

$$\hat{y}^{(ML)} = \arg \max_{y \in Y} p(y|D),$$

which maximizes the Bayesian posterior probability $p(y|D)$, is referred to as a Bayesian maximum likelihood (ML) estimator.

For problems classified as Problem 3, Bayesian ML estimators have dominated the field of estimators in bioinformatics for years. The classical solutions of Problem 1 and Problem 2 are regarded as Bayesian ML estimators with specific probability distributions, as seen in the following examples.

Example 1 (Pairwise alignment with maximum score) *In Problem 1 with a scoring model (e.g., gap costs and a substitution matrix), the distribution $p(y|D)$ in Assumption 1 is derived from the Miyazawa model [13] (See “Probability distributions $p^{(a)}(\theta|x, x')$ on $\mathcal{A}(x, x')$ ” (Section A.2.1) in Appendices), and the Bayesian ML estimator is equivalent to the alignment that has the highest similarity score.*

Example 2 (RNA structure with minimum free energy) *In Problem 2 with a McCaskill energy model [14], the distribution $p(y|D)$ in Assumption 1 can be obtained with the aid of thermodynamics (See “Probability distributions $p^{(s)}(\theta|x)$ on $\mathcal{S}(x)$ ” (Section A.2.2) in Appendices for details), and the Bayesian ML estimator is equivalent to the secondary structure that has the minimum free energy (MFE).*

When a stochastic model such as a pair hidden Markov model (pair HMM) in Problem 1 or a stochastic context-free grammar (SCFG) in Problem 2 is assumed in such problems, the distribution and the ML estimator are derived in a more direct manner.

The Bayesian ML estimator regards the solution which has the highest probability as the most likely one. To provide more general criteria for good estimators, here we define the *gain function* that gives the gain for the prediction, and the *maximum expected gain (MEG) estimator* that maximizes the *expected gain*.

Definition 2 (Gain function) *In Problem 3, for a point $\theta \in Y$ and its prediction $y \in Y$, a gain function is defined as $G : Y \times Y \rightarrow \mathbb{R}^+$, $G(\theta, y)$.*

Definition 3 (MEG estimator) *In Problem 3 with Assumption 1, the maximum expected gain (MEG) estimator is defined as*

$$\hat{y}^{(MEG)} = \arg \max_{y \in Y} \int G(\theta, y) p(\theta|D) d\theta.$$

If the gain function is designed according to the *accuracy measures* of the target problem, the MEG estimator is considered as the maximum expected accuracy (MEA) estimator, which has been successfully applied in bioinformatics (e.g., [9]). Although in estimation theory a *loss function* that should be minimized is often used, in order to facilitate the understanding of the relationship with the MEA, in this paper, we use a *gain function* that should be maximized.

The MEG estimator for the gain function $\delta(y, \theta)$ is the ML estimator. Although this means that the ML estimator maximizes the probability that the estimator is identical to the true value, there is an extensive collection of suboptimal solutions and the probability of the ML estimator is extremely small in cases where n in Problem 3 is large. Against this background, Carvalho and Lawrence proposed the *centroid estimator*, which takes into account the overall ensemble of solutions [1]. The centroid estimator can be defined as an MEG estimator for a *pointwise gain function* as follows:

Definition 4 (Pointwise gain function) *In Problem 3, for a point $\theta \in Y$ and its prediction $y = \{y_i\}_{i=1}^n \in Y$, a gain function $G(\theta, y)$ written as*

$$G(\theta, y) = \sum_{i=1}^n F_i(\theta, y_i), \quad (1)$$

where $F_i : Y \times \{0, 1\} \rightarrow \mathbb{R}^+$ ($i = 1, 2, \dots, n$), is referred to as a *pointwise gain function*.

Definition 5 (Centroid estimator [1]) In Problem 3 with Assumption 1, a centroid estimator is defined as an MEG estimator for the pointwise gain function given in Eq. (1) by defining $F_i(\theta, y_i) = I(\theta_i = 1)I(y_i = 1) + I(\theta_i = 0)I(y_i = 0)$.

Throughout this paper, $I(\cdot)$ is the indicator function that takes a value of 1 or 0 depending on whether the condition constituting its argument is true or false. The centroid estimator is equivalent to the expected Hamming loss minimizer [1]. If we can maximize the pointwise gain function independently in each dimension, we can obtain the following *consensus estimator*, which can be easily computed.

Definition 6 (Consensus estimator [1]) In Problem 3 with Assumption 1, the consensus estimator $\hat{y}^{(c)} = \{\hat{y}_i^{(c)}\}_{i=1}^n$ for a pointwise gain function is defined as

$$\hat{y}_i^{(c)} = \arg \max_{y_i \in \{0,1\}} E_{\theta|D} [F_i(\theta, y_i)] = \arg \max_{y_i \in \{0,1\}} \int F_i(\theta, y_i) p(\theta|D) d\theta.$$

The consensus estimator is generally *not* contained within the predictive space Y since the predictive space Y usually has complex constraints for each dimension (see “Discrete (binary) spaces in bioinformatics” (Section A.1) in Appendices). Carvalho and Lawrence proved a sufficient condition for the centroid estimator to contain the consensus estimator (Theorem 2 in [1]). Here, we present a more general result, namely, a sufficient condition for the MEG estimator for a pointwise function to contain the consensus estimator.

Theorem 1 In Problem 3 with Assumption 1 and a pointwise gain function, let us suppose that a predictive space Y can be written as

$$Y = \bigcap_{k=1}^K C_k, \quad (2)$$

where C_k is defined as

$$C_k = \left\{ y \in \{0,1\}^n \mid \sum_{i \in I_k} y_i \leq 1 \right\} \text{ for } k = 1, 2, \dots, K$$

for an index-set $I_k \subset \{1, 2, \dots, n\}$. If the pointwise gain function in Eq. (1) satisfies the condition

$$F_i(\theta, 1) - F_i(\theta, 0) + F_j(\theta, 1) - F_j(\theta, 0) \leq 0 \quad (3)$$

for every $\theta \in Y$ and every $i, j \in I_k$ ($1 \leq k \leq K$), then the consensus estimator is in the predictive space Y , and hence the MEG estimator contains the consensus estimator.

The above conditions are frequently satisfied in bioinformatics problems (see Supplementary Sections A.1 for examples).

3 Results

3.1 γ -centroid estimator: generalized centroid estimator

In Problem 3, the “1”s and the “0”s in the binary vector of a prediction y can be interpreted as positive and negative predictions, respectively. The respective numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for a point θ and its prediction y are denoted by $TP(\theta, y)$, $TN(\theta, y)$, $FP(\theta, y)$ and $FN(\theta, y)$, respectively (See also Eqs (S1)–(S4)).

To design a *superior* MEG estimator, it is natural to use a gain function of the following form, which yields positive scores for the number of true predictions (TP and TN) and negative scores for those of false predictions (FP and FN):

$$G(\theta, y) = \alpha_1 \text{TP}(\theta, y) + \alpha_2 \text{TN}(\theta, y) - \alpha_3 \text{FP}(\theta, y) - \alpha_4 \text{FN}(\theta, y), \quad (4)$$

where α_k is a positive constant ($k = 1, 2, 3, 4$). Note that this gain function is a pointwise gain function.

This gain function is naturally compatible with commonly used accuracy measures such as sensitivity, PPV, MCC and F-score (a function of TP, TN, FP and FN; see “Evaluation measures defined using TP, TN, FP and FN” (Section A.3) in Appendices for definitions). The following Definition 7 and Theorem 2 characterize the MEG estimator for this gain function.

Definition 7 (γ -centroid estimator) *In Problem 3 with Assumption 1 and a fixed $\gamma \geq 0$, the γ -centroid estimator is defined as the MEG estimator for the pointwise gain function given in Eq. (1) by*

$$F_i(\theta, y_i) = I(\theta_i = 0)I(y_i = 0) + \gamma I(\theta_i = 1)I(y_i = 1). \quad (5)$$

Theorem 2 *The MEG estimator for the gain function in Eq. (4) is equivalent to a γ -centroid estimator with $\gamma = \frac{\alpha_1 + \alpha_4}{\alpha_2 + \alpha_3}$.*

Theorem 2 (see Section A.6.2 for a formal proof) is derived from the following relations:

$$TP + FN = \sum_i I(\theta_i = 1) \text{ and } TN + FP = \sum_i I(\theta_i = 0).$$

The γ -centroid estimator maximizes the expected value of $TN + \gamma TP$, and includes the centroid estimator as a special case where $\gamma = 1$. The parameter γ adjusts the balance between the gain from true negatives and that from true positives.

The expected value of the gain function of the γ -centroid estimator is computed as follows (see Appendices for the derivation):

$$\sum_{i=1}^n [(\gamma + 1)p_i - 1] I(y_i = 1) + \sum_{i=1}^n (1 - p_i) \quad (6)$$

where

$$p_i = p(\theta_i = 1|D) = \sum_{\theta \in \Theta} I(\theta_i = 1)p(\theta|D). \quad (7)$$

Since the second term in Eq. (6) does not depend on y , the γ -centroid estimator maximizes the first term. The following theorem is obtained by assuming the additional condition described below.

Theorem 3 *In Problem 3 with Assumption 1, the predictive space Y satisfies the following condition: if $y = \{y_i\} \in Y$, then $y' = \{y'_i\} \in Y$ where $y'_i \in \{y_i, 0\}$ for all i . Then, the γ -centroid estimator is equivalent to the estimator that maximizes the sum of marginalized probabilities p_i that are greater than $1/(\gamma + 1)$ in the prediction.*

The condition is necessary to obtain 0 for the i that produces negative values in the first term in Eq. (6). Problem 2, Problem 1, and many other typical problems in bioinformatics satisfy this condition. Because the pointwise gain function of the γ -centroid estimator satisfies Eq. (3) in Theorem 1, we can prove the following Corollary 1.

Corollary 1 (γ -centroid estimator for $0 \leq \gamma \leq 1$) *In Problem 3 with Assumption 1, the predictive space Y is given in the same form in Eq. (2) of Theorem 1. Then, the γ -centroid estimator for $\gamma \in [0, 1]$ contains its consensus estimator. Moreover, the consensus estimator is identical to the following estimator $y^* = \{y_i^*\}$:*

$$y_i^* = \begin{cases} 1 & \text{if } p_i > \frac{1}{\gamma+1} \\ 0 & \text{if } p_i \leq \frac{1}{\gamma+1} \end{cases} \quad \text{for } i = 1, 2, \dots, n \quad (8)$$

where $p_i = p(\theta_i = 1|D) = I(\theta_i = 1)p(\theta|D)$.

Here, p_i is the marginalized probability of the distribution for the i -th dimension of the predictive space. In Problem 1, it is known as the alignment probability, which is defined as the probability of each pair of positions across the two sequences being aligned. In Problem 2, it is known as the base pairing probability, which is defined as the probability of each pair of positions forming a base pair in the secondary structure. These marginalized probabilities can be calculated by using dynamic programming algorithms, such as the forward-backward algorithm and the McCaskill algorithm, depending on the model of the distributions. (see “Probability distributions on discrete spaces” (Section A.2) in Appendices for those distributions).

Corollary 1 does not hold for $\gamma > 1$, but in typical problems in bioinformatics the γ -centroid estimator for $\gamma > 1$ can be calculated efficiently by using dynamic programming, as shown in the following examples.

Example 3 (γ -centroid estimator of pairwise alignment) *In Problem 1 with Assumption 1, the γ -centroid estimator maximizes the sum of the alignment probabilities which are greater than $1/(\gamma + 1)$ (Theorem 3), and for $\gamma \in [0, 1]$ it can be given as the consensus estimator calculated from Eq. (8) (Corollary 1). For $\gamma > 1$, the γ -centroid estimator is obtained by using a dynamic programming algorithm with the same type of iterations as in the Needleman-Wunsch algorithm:*

$$M_{i,k} = \max \begin{cases} M_{i-1,k-1} + (\gamma + 1)p_{ik} - 1 \\ M_{i-1,k} \\ M_{i,k-1} \end{cases} \quad (9)$$

where $M_{i,k}$ stores the optimal value of the alignment between two sub-sequences, $x_1 \cdots x_i$ and $x'_1 \cdots x'_k$ (see “Secondary structure prediction of an RNA sequence (Problem 2)” in Appendices for detailed descriptions).

Example 4 (γ -centroid estimator for prediction of secondary structures) *In Problem 2 with Assumption 1, the γ -centroid estimator maximizes the sum of the base pairing probabilities that are greater than $1/(\gamma + 1)$ (Theorem 3), and for $\gamma \in [0, 1]$ it can be given as the consensus estimator calculated from Eq. (8) (Corollary 1). For $\gamma > 1$, the γ -centroid estimator is obtained with the aid of a dynamic programming algorithm with the same type of iterations as in the Nussinov algorithm:*

$$M_{i,j} = \max \begin{cases} M_{i+1,j} \\ M_{i,j-1} \\ M_{i+1,j-1} + (\gamma + 1)p_{ij} - 1 \\ \max_k [M_{i,k} + M_{k+1,j}] \end{cases} \quad (10)$$

where $M_{i,j}$ stores the best score of the sub-sequence $x_i x_{i+1} \cdots x_j$ (see “Pairwise alignment of biological sequences (Problem 1)” in Appendices for the detail descriptions).

The γ -centroid estimators are implemented in LAST [4] for Problem 1 and in CentroidFold [12, 15] for Problem 2.

Problem 4 (Estimation of phylogenetic trees) *Given a set of operational taxonomic units S , predict their phylogenetic trees (unrooted and multi-branched trees) as a point in $\mathcal{T}(S)$, the space of all the possible phylogenetic trees of S .*

The phylogenetic tree in $\mathcal{T}(S)$ is represented as a binary vector with $2^{n-1} - n - 1$ dimension where n is the number of units in S , based on partition of S by cutting every edge in the tree (see “The space of phylogenetic trees: $\mathcal{T}(S)$ ” (Section A.1.3) in Appendices for details). A sampling algorithm can be used to estimate the partitioning probabilities approximately [16].

Example 5 (γ -centroid estimator of phylogenetic estimation) *In Problem 4 with Assumption 1, the γ -centroid estimator maximizes the number of the partitioning probabilities which are greater than $1/(\gamma + 1)$ (Theorem 3), and for $\gamma \in [0, 1]$ it can be give as the consensus estimator calculated from Eq. (8) (Corollary 1) (see “Estimation of phylogenetic trees (Problem 4)” in Appendices for details).*

Because the Hamming distance between two trees in $\mathcal{T}(S)$ is known as topological distance [17], the 1-centroid estimator minimizes the expected topological distance. In contrast to Example 3 and Example 4, it appears that no method can efficiently compute the γ -centroid estimator with $\gamma > 1$ in Example 5. Despite the difficulties of the application to phylogenetic trees, recently, a method applying the concept of generalized centroid estimators was developed [18].

3.2 Generalized centroid estimators for representative prediction

Predictions based on probability distributions on the predictive space were discussed in the previous sections. However, there are certain even more complex problems in bioinformatics, as illustrated by the following example.

Problem 5 (Prediction of common secondary structures of RNA sequences) *Given a set of RNA sequences $D = \{x_i\}, i = 1, \dots, K$ and their multiple alignment of length L and the same energy model for each RNA sequence, predict their common secondary structure as a point in $\mathcal{S}'(L)$, which is the space of all possible secondary structures of length L .*

In the case of Problem 5, although the probability distribution is not implemented in the predictive space, each RNA sequence x_i has a probability distribution on its secondary structure derived from the energy model. Therefore, the theories presented in the previous section cannot be applied directly to this problem. However, if we devise a new type of gain function that connects the predictive space with the parameter space of the secondary structure of each RNA sequence, we can calculate the expected gain over the distribution on the parameter spaces of RNA sequences. In order to account for this type of problem in general, we introduce Assumption 2 and Definition 8 as follows.

Assumption 2 *In Problem 3 there exists a probability distribution $p(\theta|D)$ on the parameter space Θ which might be different from the predictive space Y .*

Definition 8 (Generalized gain function) *In Problem 3 with Assumption 2, for a point $\theta \in \Theta$ and a prediction $y \in Y$, a generalized gain function is defined as $G : \Theta \times Y \rightarrow \mathbb{R}^+$, $G(\theta, y)$.*

It should be emphasized that the MEG estimator (Definition 3), pointwise gain function (Definition 4) and Theorem 1 can be extended to the generalized gain function.

In the case of Problem 5, for example, the parameter space is the product of the spaces of the secondary structures of each RNA sequence, and the probability distribution is the product of the distributions of secondary structures of each RNA sequence. Here, the general form of the problem of representative prediction is introduced.

Problem 6 (Representative prediction) *In Problem 3 with Assumption 2, if the parameter space is represented as a product space ($\Theta = \prod_{k=1}^K \Theta^{(k)} = Y^K$) and the distribution of $\theta \in \Theta$ has the form $p(\theta|D) = \prod_{k=1}^K p^{(k)}(\theta^k|D)$, predict a point y in the predictive space Y .*

The generalized gain function for the representative prediction should be chosen such that the prediction reflects as much as each data entry. Therefore, it is natural to use the following generalized gain function that integrates the gain for each parameter.

Definition 9 (Homogeneous generalized gain function) *In Problem 6, a homogeneous generalized gain function is defined as*

$$G(\theta, y) = \sum_{k=1}^K G'(\theta^k, y),$$

where G' is the gain function in Definition 2.

Definition 10 (Representative estimator) *In Problem 6, given a homogeneous generalized gain function $G(\theta, y) = \sum_{k=1}^K G'(\theta^k, y)$, the MEG estimator defined as*

$$\hat{y}^{(rMEG)} = \arg \max_{y \in Y} \int G(\theta, y) p(\theta|D) d\theta$$

is referred to as the representative estimator.

Proposition 1 *The representative estimator is equivalent to an MEG estimator with averaged probability distribution on the predictive space Y :*

$$p(y|D) = \frac{1}{K} \sum_k p^{(k)}(y|D)$$

and a gain function G' .

This proposition shows that a representative prediction problem with any homogeneous generalized gain function can be solved in a manner similar to Problem 3 ($\Theta = Y$) with averaged probability distribution. Therefore, the γ -centroid estimator for a representative prediction satisfies Corollary 2.

Corollary 2 *In Problem 6, the representative estimator where $G'(\theta^k, y)$ is the gain function of the γ -centroid estimator on Y , is the γ -centroid estimator for the averaged probability distribution and satisfies the same properties in Theorem 2, Theorem 3, and Corollary 1.*

3.3 Estimators based on marginal probabilities

In the previous section, we introduced Assumption 2, where there is a parameter space Θ that can be different from the predictive space Y , and we discussed the problem of representative prediction. In this section, we discuss another type of problems where $\Theta \neq Y$. An example is presented below.

Problem 7 (Pairwise alignment using homologous sequences) *Given a data set $D = \{x, x', h\}$, where x and x' are two biological sequences to be aligned and h is a sequence that is homologous to both x and x' , predict a point y in the predictive space $Y = \mathcal{A}(x, x')$ (the space of all possible alignments of x and x').*

The precise probabilistic model of this problem might include the phylogenetic tree, ancestor sequences and their alignments. Here, we assume a simpler situation where the probability distribution of all possible multiple alignments of D is given. We predict the pairwise alignment of two specific sequences according to the probability distribution of multiple alignments. Although the parameter space Θ , which is the space of all the possible multiple alignments, can be parametrized using the parameters of the spaces of the alignments of all pairs that can be formed from the sequences in D , Θ itself is not the product space of these spaces because these pairwise alignments are not independent: for $x, x', h \in D$, x_i must be aligned to x'_j if both x_i and x'_j are aligned to h_k . This type of problems can be generalized as follows.

Problem 8 (Prediction in a subspace of the parameter space) *In Problem 3 with Assumption 2, if the parameter space Θ is represented as $\Theta \subset \Theta' \times \Theta'^\perp$, predict a point y in the predictive space $Y = \Theta'$.*

For the problem of representative prediction (Problem 6), generalized gain functions on $\Theta \times Y$ were introduced (Definition 8 and Definition 9). In contrast, in Problem 8, the values of the parameters in Θ'^\perp are not important, and a point in $Y = \Theta'$ is predicted. In Problem 7, for example, the optimal multiple alignment of D , the pairwise alignment of h and x , and the pairwise alignment of h and x' are irrelevant, but instead we predict the pairwise alignment of x and x' . The MEG estimator for the gain function defined on $\Theta' \times Y$ can be written as

$$\hat{y}^{(sMEG)} = \arg \max_{y \in Y} \int G(\theta', y) p(\theta' | D) d\theta',$$

where $p(\theta' | D)$ on Θ' is the marginalized distribution

$$p(\theta' | D) = \int p(\theta | D) d\theta'^\perp = \int p(\theta', \theta'^\perp | D) d\theta'^\perp. \quad (11)$$

From the above MEG estimator, it might appear that Problem 8 is trivial. However, it is not a simple task to calculate the marginalized distribution in Eq. (11) in actual problems.

To reduce the computational cost, we change Problem 8 by introducing an approximated probability distribution on the product space $\Theta' \times \Theta'^\perp$ as follows.

Problem 9 (Prediction in product space) *In Problem 3 with Assumption 2, if the parameter space Θ is represented as $\Theta = \Theta' \times \Theta'^\perp$ and the probability distribution on Θ is defined as*

$$\bar{p}(\theta | D) = p(\theta' | D) p(\theta'^\perp | D), \quad (12)$$

predict a point y in the predictive space $Y = \Theta'$.

This factorization of spaces and probability distributions creates a number of inconsistencies in the parameter space with respect to the original Problem 8. In other words, the approximated distribution yields non-zero values for a point that is not included in the original Θ (in Problem 8) but in $\Theta' \times \Theta'^\perp$. To reduce these inconsistencies, a new type of gain function and a new estimator are introduced as follows.

Definition 11 (γ -type pointwise gain function) *In Problem 8, a γ -type pointwise gain function is defined as $G(\theta, y)$ in Eq. (1) in Definition 4 having*

$$F_i(\theta, y_i) = \gamma \cdot \delta_i(\theta') \cdot I(y_i = 1) + (1 - \delta_i(\theta')) I(y_i = 0), \quad (13)$$

where the value $\delta_i(\theta') \in [0, 1]$ in the gain function should be designed to reduce the inconsistencies resulting from the factorization.

Definition 12 (Approximated γ -type estimator) In Problem 9, with a γ -type pointwise gain function with $F_i(\theta, y_i)$ in Eq. (13) on $\Theta \times Y$, an approximated γ -type estimator is defined as an MEG estimator:

$$\hat{y}^{(\gamma app)} = \arg \max_{y \in Y} \int \left[\sum_{i=1}^n F_i(\theta, y_i) \right] \bar{p}(\theta|D) d\theta.$$

Example 6 (PCT in pairwise alignment) We obtain the approximate estimator for Problem 7 with the following settings. The parameter space is given as $\Theta = \Theta' \times \Theta'^\perp$, where

$$\Theta' = \mathcal{A}(x, x') (= Y) \text{ and } \Theta'^\perp = \mathcal{A}(x, h) \times \mathcal{A}(x', h)$$

and the probability distribution on the parameter space Θ is given as

$$p(\theta|D) = p^{(a)}(\theta^{xx'}|x, x') p^{(a)}(\theta^{xh}|x, h) p^{(a)}(\theta^{x'h}|x', h)$$

for $\theta = (\theta^{xx'}, \theta^{xh}, \theta^{x'h}) \in \Theta = \Theta' \times \Theta'^\perp$. The $\delta_i(\theta')$ in Eq. (13) of the γ -type pointwise gain function is defined as

$$\delta_{ik}(\theta') = \frac{1}{2} \left\{ I(\theta_{ik}^{xx'} = 1) + \sum_v I(\theta_{iv}^{xh} = 1) I(\theta_{kv}^{x'h} = 1) \right\}.$$

The approximated γ -type estimator for this γ -type pointwise gain function is employed in a part of probabilistic consistency transformation (PCT) [19], which is an important step toward accurate multiple alignments. See “Pairwise alignment using homologous sequences” (Section A.5.6) in Appendices for precise descriptions.

It is easily seen that Theorem 3 applies to the approximated γ -type estimator if p_i in Theorem 3 is changed as follows:

$$p_i = \int \delta_i(\theta') p(\theta'|D) d\theta'.$$

Moreover, to confirm whether approximated γ -type estimator contains the consensus estimator for the same gain function, it is only necessary to check if

$$(\gamma + 1) (\delta_i(\theta') + \delta_j(\theta')) - 2 \leq 0, \quad (14)$$

instead of Eq. (3) in Theorem 1. (Note that Theorem 1 can be extended to the *generalized* (pointwise) gain function: see Theorem 4.)

4 Discussion

4.1 Properties of the γ -centroid estimator

In this paper, general criteria for designing estimators are given by the maximum expected gain (MEG) estimator (Definition 3). The Bayesian ML estimator is an MEG estimator with the delta function $\delta(y, \theta)$ as the gain function, which means that only the probability for the “perfect match” is counted. To overcome the drawbacks of the Bayesian ML estimator, the centroid estimator [1] takes into account the overall ensemble of solutions and minimizes the expected Hamming loss. Because the Hamming loss is not the standard evaluation measures for actual problems, we have proposed an estimator of a more general type, the γ -centroid estimator (Definition 7), which includes the centroid estimator as a special case, $\gamma = 1$. The γ -centroid estimator is an MEG estimator that maximizes the expected value of $TN + \gamma TP$, which generally covers all possible linear combination of the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) (Theorem 2). Since most of the evaluation measures

of the prediction accuracy are functions of these numbers [20], the γ -centroid estimator is related to the principle of maximum expected accuracy (MEA). It should be noted that MEG estimators have been proposed that are similar to the γ -centroid estimator for some specific problems, for example, the alignment metric accuracy (AMA) estimator [21] (see Section A.5.1 for the formal definition) for pairwise alignment (Problem 1) and the MEA-based estimator [5] (see Appendices for the formal definition) for prediction of secondary structure of RNA (Problem 2). However, these estimators display a *bias* with respect to the accuracy measures for the problem (see Eqs. (S6) and (S8)), and are therefore inappropriate from the viewpoint of the principles of MEA. Moreover, these estimators cannot be introduced in a general setting, that is, Problem 3. It has been also shown that the γ -centroid estimator outperforms the MEA-based estimator [5] for various probability distributions in computational experiments [12]. (See “Pairwise alignment of biological sequences (Problem 1)” and “Secondary structure prediction of an RNA sequence (Problem 2)” in Appendices for relations between the γ -centroid estimator and other estimators in Problems 1 and 2, respectively.)

4.2 How to determine the parameter in γ -centroid estimator

The parameter γ in γ -centroid estimators adjusts sensitivity and PPV (whose relation is trade-off). MCC or F-score is often used to obtain a balanced measure between sensitivity and PPV. In RNA secondary structure predictions, it has been confirmed that the best γ (with respect to MCC) of the γ -centroid estimator with CONTRAfold model was larger than that with McCaskill model [12]. It shows that the best γ (with respect to a given accuracy measure) depends on not only estimation problems but also probabilistic models for predictive space. The parameter γ trained by using reference structures was therefore employed as the default parameter in CentroidFold [12]. In order to select the parameter automatically (with respect to a given accuracy measure such as MCC and F-score), an approximation of maximizing expected MCC (or F-score) with the γ -centroid estimator can be utilized [22].

4.3 Accuracy measures and computational efficiency

The reader might consider that it is possible to design estimators that maximize the expected MCC or F-score which balances sensitivity (SEN) and positive predictive value (PPV). However, it is much more difficult to compute such estimators in comparison with the γ -centroid estimator, as described below.

The expected value of the gain function of the γ -centroid estimator can be written with marginalized probabilities as in Eq. (7), which can be efficiently computed by dynamic programming in many problems in bioinformatics, for example, the forward-backward algorithm for alignment probabilities and the McCaskill algorithm for base pairing probabilities. Under a certain condition of the predictive space, which many problems in bioinformatics satisfy, the γ -centroid estimator maximizes the sum of marginalized probabilities greater than $1/(\gamma + 1)$ (Theorem 3). Moreover, under an additional condition of the predictive space and the pointwise gain function, which again many problems in bioinformatics satisfy, the γ -centroid estimators for $\gamma \in [0, 1]$ can be easily calculated as the consensus estimators, which collect in the binary predictive space the components that have marginalized probabilities greater than $1/(\gamma + 1)$ (Corollary 1). For $\gamma > 1$, there often exist dynamic programming algorithms that can efficiently compute the γ -centroid estimators (Examples 4 & 3), but there are certain problems, such as Problem 4, which seem to have no efficient dynamic programming algorithms.

The gain function of the estimators that maximize MCC or F-score, and also SEN or PPV contain *multiplication* and/or *division* of TP, TN, FP and FN, while the gain function of the γ -centroid estimator contains only the weighted *sums* of these values (i.e., $TN + \gamma \cdot TP$). Therefore, the expected gain is not written with marginalized probabilities as in Eq. (7), and it is difficult to design efficient computational algorithms for those estimators. In predicting secondary structures of RNA sequences (Problem 2), for example, it is necessary to enumerate all candidate

secondary structures or sample secondary structures for an approximation in order to compute the expected MCC/F-score of a predicted secondary structure.

4.4 Probability distributions are not always defined on predictive space

After discussing the standard estimation problems on a binary space where the probability distribution is defined on the predictive space, we have proposed a new category of estimation problems where the probability distribution is defined on a parameter space that differs from the predictive space (see Assumption 2). Two types of estimators for such problems, for example, estimators for representative prediction and estimators based on marginalized distribution, have been discussed.

Prediction of the common secondary structure from an alignment of RNA sequences (Problem 5) is an example of representative prediction. The probability distribution is not implemented in the predictive space, the space of common secondary structure, but each RNA sequence has a probability distribution for its secondary structure. Because the “correct” reference for the common secondary structure is not known in general, direct evaluation of the estimated common secondary structure is difficult. In the popular evaluation process for this problem, the predicted common secondary structure is mapped to each RNA sequence and compared to its reference structure. Using the homogeneous generalized gain function exactly implements this evaluation process and the MEG estimator for the averaged probability distribution is equivalent to the MEG estimator for homogeneous generalized gain function. Therefore, we can use the averaged base pairing probabilities according to the alignment as the distribution for the common secondary structure (see “Common secondary structure prediction from a multiple alignment of RNA sequences” (Section A.5.5) in Appendices for detailed discussion). The representative estimator for Problem 5 is implemented in software **CentroidAlifold**. Another example of representative prediction is the “alignment of alignments” problem, which is the fundamental element of progressive multiple alignment of biological sequences. The evaluation process using the sum of pairs score corresponds to using the homogeneous generalized gain function. (see “Alignment between two *alignments* of biological sequences” (Section A.5.4) in Appendices for detailed discussion).

Estimation problems of marginalized distributions can be formalized as prediction in a subspace of the parameter space (Problem 8). If we can calculate the marginalized distribution on the predictive space from the distribution on the parameter space, all general theories apply to the predictive space and the marginalized distribution. In actual problems, such as pairwise alignment using homologous sequences (Problem 7), however, computational cost for calculation of the marginalized probability is quite high. We introduced the factorized probability distribution (Eq. (12)) for approximation, the γ -type pointwise gain function (Definition 11) to reduce the inconsistency caused by the factorization, and the approximated γ -type estimator (Definition 12). In Problem 7, the probability consistency transformation (PCT), which is widely used for multiple sequence alignment, is interpreted as an approximated γ -type estimator. Prediction of secondary structures of RNA sequences on the basis of homologous sequences [23] (see Problem 13 in Appendices) and pairwise alignment for *structured* RNA sequences are further examples of this type of problems.

4.5 Application of γ -centroid estimator to cluster centroid

In case probability distribution on the predictive space is multi-modal, γ -centroid estimators can provide unreliable solutions. For example, when there are two clusters of secondary structures in predictive spaces and those structures are exclusive, the γ -centroid estimator might give a “chimeric” secondary structure whose free energy is quite high. To avoid this situation, Ding *et al.* [24] proposed a notion of the *cluster centroid*, which is computed by the centroid estimator with a given cluster in a predictive space. We emphasize that the extension of cluster centroid by using γ -centroid estimator is straightforward and would be useful.

4.6 Conclusion

In this work, we constructed a general framework for designing estimators for estimation problems in high-dimensional discrete (binary) spaces. The theory is regarded as a generalization of the pioneering work conducted by Carvalho and Lawrence, and is closely related to the concept of MEA. Furthermore, we presented several applications of the proposed estimators (see Table 1 for summary) and the underlying theory. The concept presented in this paper is highly extendable and sheds new light on many problems in bioinformatics. In future research, we plan to investigate further applications of the γ -centroid and related estimators presented in this paper.

Acknowledgments

The authors are grateful to Drs. Luis E. Carvalho, Charles E. Lawrence, Kengo Sato, Toutai Mituyama and Martin C. Frith for fruitful discussions. The authors also thank the members of the bioinformatics group for RNA at the National Institute of Advanced Industrial Science and Technology (AIST) for useful discussions.

Table 1: Summary of applications in bioinformatics

Alignment	(1) Pairwise alignment of biological sequences	(4) Pairwise alignment of two multiple alignments	(6) Pairwise alignment using homologous sequences	
Section	Section A.5.1	Section A.5.4	Section A.5.6	
Data D	$\{x, x'\}$	$\{A, A'\}$	$\{x, x', H\}$	
Predictive space Y	$\mathcal{A}(x, x')$	$\mathcal{A}(A, A')$	$\mathcal{A}(x, x')$	
Parameter space Θ	$\mathcal{A}(x, x')$	$\prod_{x \in A} \prod_{x' \in A'} \mathcal{A}(x, x')$	$\mathcal{A}(x, x') \times \prod_{h \in H} [\mathcal{A}(x, h) \times \mathcal{A}(x', h)]$	
Probability $p(\theta D)$	$p^{(a)}(\theta x, x')$	$\prod_{x \in A} \prod_{x' \in A'} p^{(a)}(\theta x, x')$	$p^{(a)}(\theta^{xx'} x, x') \prod_{h \in H} [p^{(a)}(\theta^{xh} x, h) p^{(a)}(\theta^{x'h} x', h)]$	
Type of estimator	γ -centroid	representative	approximate	
Software	LAST	—	—	
Reference	[4]	[19], This work	[19], This work	
RNA	(2) Secondary structure prediction of RNA	(5) Common secondary structure prediction	(7) Secondary structure prediction using homologous sequences	(8) Pairwise alignment of structured RNAs
Section	Section A.5.2	Section A.5.5	Section A.5.7	Section A.5.8
Data D	$\{x\}$	$\{A\}$	$\{x, H\}$	$\{x, x'\}$
Predictive space Y	$\mathcal{S}(x)$	$\mathcal{S}(A)$	$\mathcal{S}(x)$	$\mathcal{A}(x, x')$
Parameter space Θ	$\mathcal{S}(x)$	$\prod_{x \in A} \mathcal{S}(x)$	$\mathcal{S}(x) \times \prod_{h \in H} [\mathcal{A}(x, h) \times \mathcal{S}(h)]$	$\mathcal{A}(x, x') \times \mathcal{S}(x) \times \mathcal{S}(x')$
Probability $p(\theta D)$	$p^{(s)}(\theta x)$	$\prod_{x \in A} p^{(s)}(\theta x)$	$p^{(s)}(\theta^x x) \times \prod_{h \in D} [p^{(a)}(\theta^{xh} x, h) p^{(s)}(\theta^h h)]$	$p^{(a)}(\theta^{xx'} x, x') p^{(s)}(\theta^x x) p^{(s)}(\theta^{x'} x')$
Type of estimator	γ -centroid	representative	approximate	approximate
Software	CENTROIDFOLD	CENTROIDALIFOLD	CENTROIDHOMFOLD	CENTROIDALIGN
Reference	[12]	[12, 25]	[23]	[26]
Phylogenetic tree	(3) Estimation of phylogenetic tree			
Section	Section A.5.3			
Data D	S			
Parameter space Θ	$\mathcal{T}(S)$			
Predictive space Y	$\mathcal{T}(S)$			
Probability $p(\theta D)$	$p^{(t)}(\theta S)$			
Type of estimator	γ -centroid			
Reference	This work			

The top row includes problems about RNA secondary structure predictions and the middle row includes problems about alignment of biological sequences. Note that the estimators in the same column corresponds to each other.

A Appendices

A.1 Discrete (binary) spaces in bioinformatics

In this section, we summarize three discrete spaces that appear in this paper. These discrete spaces are often used in the definition of the predictive spaces and the parameter spaces. It should be noted that every discrete space described below is identical in form to Eq. (2).

A.1.1 The space of alignments of two biological sequences: $\mathcal{A}(x, x')$

We define a space of the alignments of two biological (DNA, RNA and protein) sequences x and x' , denoted by $\mathcal{A}(x, x')$, as follows. We set $I^{(0)} = \{(i, k) | 1 \leq i \leq |x|, 1 \leq k \leq |x'|\}$ as a base index set, and a binary variable θ_{ik} for $(i, k) \in I^{(0)}$ is defined by

$$\theta_{ik} = \begin{cases} 1 & \text{positions } i \text{ in } x \text{ and } k \text{ in } x' \text{ are aligned} \\ 0 & \text{positions } i \text{ in } x \text{ and } k \text{ in } x' \text{ are not aligned} \end{cases}.$$

Then $\mathcal{A}(x, x')$ is a subset of $B := \left\{ \theta = \{\theta_{ik}\}_{(i,k) \in I^{(0)}} \mid \theta_{ik} \in \{0, 1\} \right\}$ and is defined by

$$\mathcal{A}(x, x') = \bigcap_{I \in \mathcal{I}} C(I), \quad C(I) = \left\{ x' \in B \mid \sum_{(i,k) \in I} \theta_{ik} \leq 1 \right\}.$$

Here \mathcal{I} is a set of index-sets:

$$\mathcal{I} = \left\{ I \mid I = I_i^{(1)} \ (1 \leq i \leq |x|) \text{ or } I = I_k^{(2)} \ (1 \leq k \leq |x'|) \text{ or } I = I_{ikjl}^{(3)} \ (1 \leq i < j \leq |x|, 1 \leq l < k \leq |x'|) \right\}$$

where

$$I_i^{(1)} = \{(i, k) | 1 \leq k \leq |x'|\}, I_k^{(2)} = \{(i, k) | 1 \leq i \leq |x|\} \text{ and } I_{ikjl}^{(3)} = \{(i, k), (j, l)\}.$$

The inclusion $y \in C(I_i^{(1)})$ means that position i in the sequence x aligns with *at most one* position in the sequence x' in the alignment y , $y \in C(I_j^{(2)})$ means that position j in the sequence x' aligns with *at most one* position in the sequence x and $y \in C(I_{ikjl}^{(3)})$ means the alignment (i, k) and (j, l) is *not crossing*. Note that $\mathcal{A}(x, x')$ depends on only the length of two sequences, namely, $|x|$ and $|x'|$.

A.1.2 The space of secondary structures of RNA: $\mathcal{S}(x)$

We define a space of the secondary structures of an RNA sequence x , denoted by $\mathcal{S}(x)$, as follows. We set $I^{(0)} = \{(i, j) | 1 \leq i < j \leq |x|\}$ as a base index set, and a binary variable θ_{ij} for $(i, j) \in I^{(0)}$ is defined by

$$\theta_{ij} = \begin{cases} 1 & \text{the positions } i \text{ of } x \text{ and } j \text{ of } x \text{ form a base pair} \\ 0 & \text{the positions } i \text{ of } x \text{ and } j \text{ of } x \text{ do not form a base pair} \end{cases}.$$

Then $\mathcal{S}(x)$ is a subset of $B := \left\{ \theta = \{\theta_{ij}\}_{(i,j) \in I^{(0)}} \mid \theta_{ij} \in \{0, 1\} \right\}$ and is defined by

$$\mathcal{S}(x) = \bigcap_{I \in \mathcal{I}} C(I), \quad C(I) = \left\{ \theta \in B \mid \sum_{(i,j) \in I} \theta_{ij} \leq 1 \right\}.$$

Here \mathcal{I} is a set of index-sets

$$\mathcal{I} = \left\{ I \mid I = I_i^{(1)} \ (1 \leq i \leq |x|) \text{ or } I = I_{ijkl}^{(2)} \ (1 \leq i < k < j < l \leq |x|) \right\}$$

where

$$I_i^{(1)} = \{(i, j) | i < j \leq |x|\} \cup \{(j, i) | 1 \leq j < i\} \text{ and } I_{ijkl}^{(2)} = \{(i, j), (k, l)\}.$$

The inclusion $y \in C(I_i^{(1)})$ means that position i in the sequence x belongs to *at most one* base-pair in a secondary structure y , and $y \in C(I_{ijkl}^{(2)})$ means two base-pairs whose relation is *pseudo-knot* are not allowed in y . Note that $\mathcal{S}(x)$ depends on only the length of the RNA sequence x , that is, $|x|$.

A.1.3 The space of phylogenetic trees: $\mathcal{T}(S)$

We define a space of phylogenetic trees (unrooted and multi-branch trees) of a set of $S = \{1, \dots, n\}$, denoted by $\mathcal{T}(S)$, as follows. We set $I^{(0)} = \{X|X \subset S^2, |X| < n/2 \vee (|X| = n/2 \wedge 1 \in X)\}$, where $S^2 = \{X|X \subset S, |X| > 1 \wedge |X| < n-1\}$, as a base index set and we define binary variables θ_X for $X \in I^{(0)}$ by

$$\theta_X = \begin{cases} 1 & \text{if } S \text{ can be partitioned into } X \text{ and } S \setminus X \text{ by cutting an edge in the tree} \\ 0 & \text{otherwise} \end{cases}.$$

Then $\mathcal{T}(S)$ is a subset of $B := \left\{ \theta = \{\theta_X\}_{X \in I^{(0)}} \mid \theta_X \in \{0, 1\} \right\}$ and is defined by

$$\mathcal{T}(S) = \bigcap_{I \in \mathcal{I}} C(I), \quad C(I) = \left\{ \theta \in B \mid \sum_{X \in I} \theta_X \leq 1 \right\}$$

where $\mathcal{I} = \{I = \{X, Y\} \mid X \cap Y \notin \{\emptyset, X, Y\}\}$. Note that $\mathcal{T}(S)$ depends on only the number of elements in S . We now give several properties of $\mathcal{T}(S)$ that follow directly from the definition.

Lemma 1 *The number of elements in $\mathcal{T}(S)$ (i.e. $|I^{(0)}|$) is equal to $2^{n-1} - n - 1$ where $n = |S|$.*

Lemma 2 *The topological distance [17] between two phylogenetic trees T_1 and T_2 in $\mathcal{T}(S)$ is*

$$d(T_1, T_2) = \sum_{X \in I^{(0)}} I(\theta_X(T_1) \neq \theta_X(T_2))$$

where $I(\cdot)$ is the indicator function.

Remark 1 *If we assume the additional condition $\sum_X \theta_X = ((4n-6) - 2n)/2 = n-3$, then $\mathcal{T}(S)$ is a set of binary trees.*

A.2 Probability distributions on discrete spaces

We use three probability distributions in this paper.

A.2.1 Probability distributions $p^{(a)}(\theta|x, x')$ on $\mathcal{A}(x, x')$

For two protein sequences x and x' , a probability distribution $p^{(a)}(\theta|x, x')$ over the space $\mathcal{A}(x, x')$, which is the space of pairwise alignments of x and x' defined in the previous section, is given by the following models.

1. Miyazawa model [13] and Probalign model [27]:

$$p^{(a)}(\theta|x, x') = \frac{1}{Z(T)} \exp\left(\frac{S(\theta)}{T}\right)$$

where $S(\theta)$ is the score of an alignment θ under the given scoring matrix (We define $S(\theta) = \sum_{\theta_{ij}=1} s(x_i, x_j) - (\text{penalty for gaps})$ where $s(x_i, x_j)$ is a score for the correspondence of bases x_i and x_j), T is the thermodynamic temperature and $Z(T)$ is the normalization constant, which is known as a *partition function*.

2. Pair Hidden Markov Model (pair HMM) [19]:

$$p^{(a)}(\theta|x, x') = \pi(s_1) \left(\prod_{i=1}^{n-1} \alpha(s_i \rightarrow s_{i+1}) \right) \left(\prod_{i=1}^n \beta(o_i|s_i) \right)$$

where $\pi(s)$ is the initial probability of starting in state s , $\alpha(s_i \rightarrow s_{i+1})$ is the transition probability from s_i to s_{i+1} and $\beta(o_i|s_i)$ is the omission probability for either a single letter or aligner residue pair o_i in the state s_i .

3. CONTRAlign (pair CRF) model [28]:

$$p^{(a)}(\theta|x, x') = \frac{\exp(w^t f(\theta, x, x'))}{\sum_{\theta' \in \Omega(x, x')} \exp(w^t f(\theta', x, x'))}$$

where w is a parameter vector and $f(\theta, x, x')$ is a vector of features that indicates the number of times each parameter appears, $\Omega(x, x')$ denotes the set of all possible alignments of x and x' . We do not describe the feature vectors and refer readers to the original paper [28].

Remark 2 *Strictly speaking, the alignment space in the pair hidden Markov model and the CONTRAlign model consider the patterns of gaps. In these cases, we obtain the probability space on $\mathcal{A}(x, x')$ by a marginalization.*

A.2.2 Probability distributions $p^{(s)}(\theta|x)$ on $\mathcal{S}(x)$

For an RNA sequence x , a probability distribution $p^{(s)}(\theta|x)$ over $\mathcal{S}(x)$, which is the space of secondary structures of x defined in the previous section is given by the following models.

1. McCaskill model [14]: This model is based on the energy models for secondary structures of RNA sequences and is defined by

$$p^{(s)}(\theta|x) = \frac{1}{Z(x)} \exp\left(-\frac{E(\theta, x)}{kT}\right) \text{ where } Z(x) = \sum_{\theta \in \mathcal{S}(x)} \exp\left(-\frac{E(\theta, x)}{kT}\right)$$

where $E(\theta, x)$ denotes the energy of the secondary structure that is computed using the energy parameters of Turner Lab [29], k and T are constants and $Z(x)$ is the normalization term known as the *partition function*.

2. Stochastic Context free grammars (SCFGs) model [30]:

$$p^{(s)}(\theta|x) = \frac{\sum_{\sigma \in \Omega(\theta)} p(x, \sigma)}{\sum_{\sigma \in \Omega'(x)} p(x, \sigma)}$$

where $p(x, \sigma)$ is the joint probability of generating the parse σ and is given by the product of the transition and emission probabilities of the SCFG model and $\Omega'(x)$ is all parses of x , $\Omega(\theta)$ is all parses for a given θ .

3. CONTRAfold (CRFs; conditional random fields) model [5]: This model gives us the best performance on secondary structure prediction although it is not based on the energy model.

$$p^{(s)}(\theta|x) = \frac{\sum_{\sigma \in \Omega(\theta)} \exp(w^t f(x, \sigma))}{\sum_{\sigma \in \Omega'(x)} \exp(w^t f(x, \sigma))}$$

where $w \in \mathbb{R}^n$, $f(x, \sigma) \in \mathbb{R}^n$ is the feature vector for x in parse σ , $\Omega'(x)$ is all parses of x , $\Omega(\theta)$ is all parses for a given θ .

A.2.3 Probability distributions $p^{(t)}(\theta|S)$ on $\mathcal{T}(S)$

A probability distribution $p^{(t)}(\theta|S)$ on $\mathcal{T}(S)$ is given by probabilistic models of phylogenetic trees, for example, [31, 32]. Those models give a probability distribution on binary trees and we should marginalize these distributions for multi-branch trees.

A.3 Evaluation measures defined using TP, TN, FP and FN

There are several evaluation measures of a prediction in estimation problems for which we have a reference (correct) prediction in Problem 3. The Sensitivity (SEN), Positive Predictive Value (PPV), Matthew's correlation coefficient (MCC) and F-score for a prediction are defined as follows.

$$\begin{aligned} \text{SEN} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{PPV} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \\ \text{F-score} &= \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \end{aligned}$$

where TP (the number of true positive), TN (the number of true negative), FP (the number of false positive) and FN (the number of false negative) are defined by

$$TP = TP(\theta, y) = \sum_i I(y_i = 1)I(\theta_i = 1), \quad (S1)$$

$$TN = TN(\theta, y) = \sum_i I(y_i = 0)I(\theta_i = 0), \quad (S2)$$

$$FP = FP(\theta, y) = \sum_i I(y_i = 1)I(\theta_i = 0), \quad (S3)$$

$$FN = FN(\theta, y) = \sum_i I(y_i = 0)I(\theta_i = 1). \quad (S4)$$

It should be noted that these measures can be written as a function of TP, TN, FP and FN. See [20] for other evaluation measures.

A.4 Schematic diagrams of representative and approximated γ -type estimators

The schematic diagrams of the MEG estimator (Definition 3), the representative estimator (Definition 10) and the approximated γ -type estimator (Definition 12) are shown in Figure S1, Figure S2 and Figure S3, respectively.

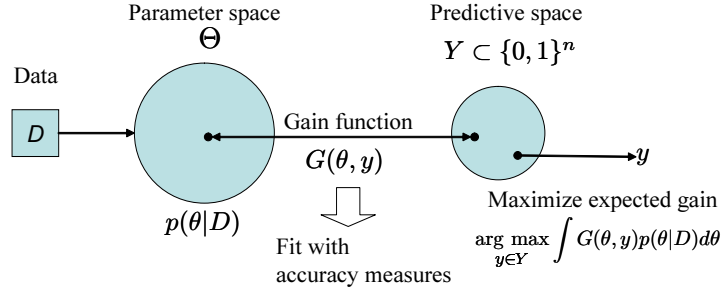


Figure S1: Schematic diagram of the MEG estimator (Definition 3).

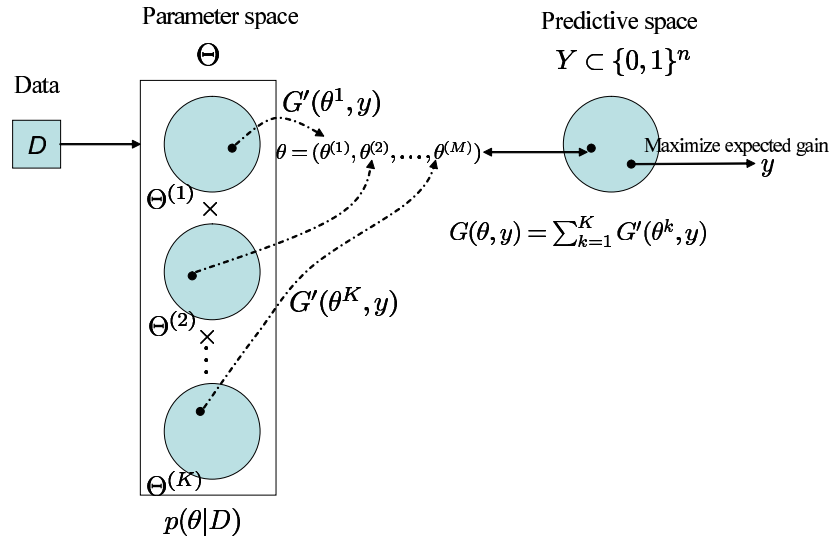


Figure S2: Schematic diagram of the representative estimator (Definition 10). The parameter space Θ is a product space and is different from the predictive space Y .

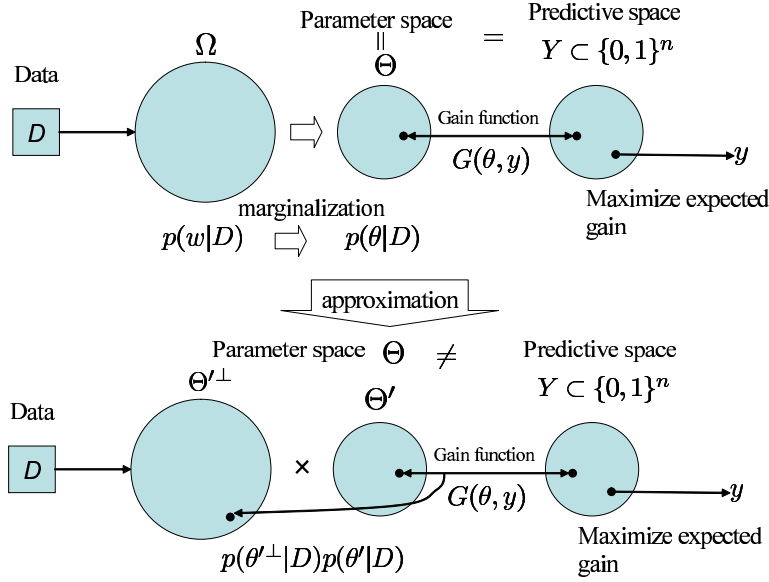


Figure S3: **Schematic diagram of the approximated γ -type estimator (Definition 12).** The estimator in the top figure shows the γ -centroid estimator with the marginalized probability distribution, and the one in the bottom figure shows its approximation.

A.5 Applications in bioinformatics

In this section we describe several applications to bioinformatics of the general theories. Some of these applications have already been published. In those cases, we briefly explain the applications and the readers should see the original paper for further descriptions as well as the computational experiments. All of the applications in this section are summarized in Table 1.

A.5.1 Pairwise alignment of biological sequences (Problem 1)

The pairwise alignment of biological (DNA, RNA, protein) sequences (Problem 1) is another fundamental and important problem of sequence analysis in bioinformatics (cf. [33]).

The γ -centroid estimator for Problem 1 can be introduced as follows:

Estimator 1 (γ -centroid estimator for Problem 1) *For Problem 1, we obtain the γ -centroid estimator where the predictive space Y is equal to $\mathcal{A}(x, x')$ and the probability distribution on Y is taken by $p^{(a)}(\theta|x, x')$.*

First, Theorem 2 and the definition of $\mathcal{A}(x, x')$ lead to the following property.

Property 1 (A relation of Estimator 1 with accuracy measures) *The γ -centroid estimator for Problem 1 is suitable for the accuracy measures: SEN, PPV, MCC and F-score with respect to the aligned-bases in the predicted alignment.*

Note that accurate prediction of aligned-bases is important for the analysis of alignments, for example, in phylogenetic analysis. Therefore, the measures in above are often used in evaluations of alignments e.g. [4].

The marginalized probability $p_{ik} = p^{(a)}(\theta_{ik} = 1|x, x') = \sum_{\theta \in \mathcal{A}(x, x')} I(\theta_{ik} = 1)p^{(a)}(\theta|x, x')$ is called the *aligned-base (matching) probability* in this paper. The aligned-base probability matrix $\{p_{ik}\}_{i,k}$ can be computed by the forward-backward algorithm whose time complexity is equal to $O(|x||x'|)$ [33]. Now, Theorem 3 leads to the following property.

Property 2 (Computation of Estimator 1) *The pairwise alignment of Estimator 1 is found by maximizing the sum of aligned-base probabilities p_{ik} (of the aligned-bases in the predicted alignment) that are larger than $1/(\gamma+1)$. Therefore, it can be computed by a Needleman-Wunsch-style dynamic programming*

(DP) algorithm [34] after calculating the aligned-base matrix $\{p_{ik}\}$:

$$M_{i,k} = \max \begin{cases} M_{i-1,k-1} + (\gamma + 1)p_{ik} - 1 \\ M_{i-1,k} \\ M_{i,k-1} \end{cases} \quad (\text{S5})$$

where $M_{i,k}$ stores the optimal value of the alignment between two sub-sequences, $x_1 \cdots x_i$ and $x'_1 \cdots x_k$.

The time complexity of the recursion of the DP algorithm in Eq. (S5) is equal to $O(|x||x'|)$, so the total computational cost for predicting the secondary structure of the γ -centroid estimator remains $O(|x||x'|)$.

By using Corollary 1, we can predict the pairwise alignment of Estimator 1 with $\gamma \in [0, 1]$ without using the DP algorithm in Eq. (S5).

Property 3 (Computation of Estimator 1 with $0 \leq \gamma \leq 1$) *The pairwise alignment of the γ -centroid estimator can be predicted by collecting the aligned-bases whose probabilities are larger than $1/(\gamma + 1)$.*

The genome alignment software called LAST (<http://last.cbrc.jp/>) [4, 35] employs the γ -centroid estimator accelerated by an X-drop algorithm, and the authors indicated that Estimator 1 reduced the false-positive aligned-bases, compared to the conventional alignment (maximum score estimator).

Relations of Estimator 1 with existing estimators are summarized as follows:

1. A relation with the estimator by Miyazawa [13] (i.e. the centroid estimator):
Estimator 1 where $\gamma = 1$ and the Miyazawa model is equivalent to the centroid estimator proposed by Miyazawa [13].
2. A relation with the estimator by Holmes *et al.* [36]:
Estimator 1 with sufficiently large γ is equivalent to the estimator proposed by Holmes *et al.*, which maximizes the sum of matching probabilities in the predicted alignment.
3. A relation with the estimator in PROBCONS: In the program, PROBCONS, Estimator 1 with pair HMM model and the sufficient large γ was used. This means that PROBCONS only take care the sensitivity (or SPS) for the predicted alignment.
4. A relation with the estimator by Schwartz *et al.*:

For Problem 1, Schwartz *et al.* [21] proposed an Alignment Metric Accuracy (AMA) estimator, which is similar to the γ -centroid estimator (see also [3]). The AMA estimator is a maximum gain estimator (Definition 3) with the following gain function.

$$G^{(\text{AMA})}(\theta, y) = 2 \sum_{i,j} I(\theta_{ij} = 1)I(y_{ij} = 1) + G_f \left\{ \sum_i \prod_j I(\theta_{ij} = 0)I(y_{ij} = 0) + \sum_j \prod_i I(\theta_{ij} = 0)I(y_{ij} = 0) \right\}$$

for $\theta, y \in \mathcal{A}(x, x')$. In the above equation, $G_f \geq 0$ is a gap factor, which is a weight for the prediction of gaps. We refer to the function $G^{(\text{AMA})}(\theta, y)$ as the gain function of the AMA estimator. In a similar way to that described in the previous section, we obtain a relation between $G^{(\text{AMA})}(\theta, y)$ and $G^{(\text{centroid})}(\theta, y)$ (the gain function of the γ -centroid estimator). If we set $1/G_f = \gamma$, then we obtain

$$G^{(\text{AMA})}(\theta, y) = \frac{2}{\gamma} G^{(\text{centroid})}(\theta, y) + \frac{1}{\gamma} A(\theta, y) + C_\theta \quad (\text{S6})$$

where

$$A(\theta, y) = \sum_i \sum_{(j_1, j_2): j_1 \neq j_2} I(\theta_{ij_1} = 1)I(y_{ij_2} = 1) + \sum_j \sum_{(i_1, i_2): i_1 \neq i_2} I(\theta_{i_1 j} = 1)I(y_{i_2 j} = 1)$$

and C_θ is a value which does not depend on y . 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 $I(\theta_{ij_2} = 0)I(y_{ij_2} = 1) = 1$, and this means that (i, j_1) is an aligned pair that is a false negative and (i, j_2) is an aligned pair that is a false positive when θ is a reference alignment and y is a predicted alignment. Therefore, the terms $A(\theta, y)$ (in Eq. (S6)) in the gain function of AMA are not appropriate for the evaluation measures SEN, PPV, MCC and F-score for aligned bases. In summary, the γ -centroid estimator is suitable for the evaluation measures: SEN, PPV and F-score with respect to the aligned-bases while the AMA estimator is suitable for the AMA.

A.5.2 Secondary structure prediction of an RNA sequence (Problem 2)

Secondary structure prediction of an RNA sequence (Problem 2) is one of the most important problems of sequence analysis in bioinformatics. Its importance has increased due to the recent discovery of functional non-coding RNAs (ncRNAs) because the functions of ncRNAs are closely related to their secondary structures [37].

γ -centroid estimator for Problem 2 can be introduced as follows:

Estimator 2 (γ -centroid estimator for Problem 2) *For Problem 2, we obtain the γ -centroid estimator (Definition 7) where the predictive space Y is equal to $\mathcal{S}(x)$ and the probability distribution on Y is taken by $p^{(s)}(\theta|x)$.*

The general theory of the γ -centroid estimator leads to several properties. First, the following property is derived from Theorem 2 and the definition of $\mathcal{S}(x)$.

Property 4 (A relation of Estimator 2 with accuracy measures) *The γ -centroid estimator for Problem 2 is suitable for the widely-used accuracy measures of the RNA secondary structure prediction: SEN, PPV and MCC with respect to base-pairs in the predicted secondary structure.*

Because the base-pairs in a secondary structure are biologically important, SEN, PPV and MCC with respect to base-pairs are widely used in evaluations of RNA secondary structure prediction, for example, [5, 12, 38].

The marginalized probability $p_{ij} = p^{(s)}(\theta_{ij} = 1|x) = \sum_{\theta \in \mathcal{S}(x)} I(\theta_{ij} = 1)p^{(s)}(\theta|x)$ is called a *base-pairing probability*. The base-pairing probability matrix $\{p_{ij}\}_{i < j}$ can be computed by the Inside-Outside algorithm whose time complexity is equal to $O(|x|^3)$ where $|x|$ is the length of RNA sequence x [14, 33]. Then, Theorem 3 leads to the following property.

Property 5 (Computation of Estimator 2) *The secondary structure of Estimator 2 is found by maximizing the sum of the base-pairing probabilities p_{ij} (of the base-pairs in the predicted structure) that are larger than $1/(\gamma + 1)$. Therefore, it can be computed by a Nussinov-style dynamic programming (DP) algorithm [39] after calculating the base-pairing probability matrix $\{p_{ij}\}$:*

$$M_{i,j} = \max \begin{cases} M_{i+1,j} \\ M_{i,j-1} \\ M_{i+1,j-1} + (\gamma + 1)p_{ij} - 1 \\ \max_k [M_{i,k} + M_{k+1,j}] \end{cases} \quad (S7)$$

where $M_{i,j}$ stores the best score of the sub-sequence $x_i x_{i+1} \cdots x_j$.

If we replace “ $(\gamma+1)p_{ij}-1$ ” with “1” in Eq. (S7), the DP algorithm is equivalent to the Nussinov algorithm [39] that maximizes the number of base-pairs in a predicted secondary structure. The time complexity of the recursion of the DP algorithm in Eq. (S7) is equal to $O(|x|^3)$. Hence, the total computational cost for predicting the secondary structure of the γ -centroid estimator remains $O(|x|^3)$, which is the same time complexity as for standard software: MFOLD [40], RNAFOLD [41] and RNASTRUCTURE [42].

By using Corollary 1, we can predict the secondary structure of Estimator 2 with $\gamma \in [0, 1]$ without using the DP algorithm in Eq. (S7).

Property 6 (Computation of Estimator 2 with $0 < \gamma \leq 1$) *The secondary structure of the γ -centroid estimator with $\gamma \in [0, 1]$ can be predicted by collecting the base-pairs whose probabilities are larger than $1/(\gamma + 1)$.*

The software CENTROIDFOLD [12, 15] implements Estimator 2 with various probability distributions for the secondary structures, such as the CONTRAFOLD and MCCASKILL models.

Relations of Estimator 2 with other estimators are summarized as follows:

1. A relation with the estimator used in SFOLD [43, 44]:

Estimator 2 with $\gamma = 1$ and the McCaskill model (i.e. the centroid estimator with the McCaskill model) is equivalent to the estimator used in the SFOLD program.

2. A relation with the estimator used in CONTRAFOLD:

For Problem 2, Do *et al.* [5] proposed an MEA-based estimator, which is similar to the γ -centroid estimator. (The MEA-based estimator was also used in a recent paper [6].) The MEA-based

estimator is defined by the maximum expected gain estimator (Definition 3) with the following gain function for θ and $y \in \mathcal{S}(x)$.

$$G^{(\text{contra})}(\theta, y) = \sum_{i=1}^{|x|} \left[\gamma \sum_{j:j \neq i} I(\theta_{ij}^* = 1) I(y_{ij}^* = 1) + \prod_{j:j \neq i} I(\theta_{ij}^* = 0) I(y_{ij}^* = 0) \right] \quad (\text{S8})$$

where θ^* and y^* are symmetric extensions of (upper triangular matrices) θ and y , respectively (i.e. $\theta_{ij}^* = \theta_{ij}$ for $i < j$ and $\theta_{ij}^* = \theta_{ji}$ for $j < i$; the definition of y^* is similar.). It should be noted that, under the general estimation problem of Problem 3, the gain function of Eq. (S8) cannot be introduced, and the gain function is specialized for the problem of RNA secondary structure prediction.

The relation between the gain function of the γ -centroid estimator (denoted by $G^{(\text{centroid})}(\theta, y)$ and defined in Definition 7) and the one of the MEA-based estimator is

$$G^{(\text{contra})}(\theta, y) = G^{(\text{centroid})}(\theta, y) + A(\theta, y) + C(\theta) \quad (\text{S9})$$

where the additional term $A(\theta, y)$ is positive for *false* predictions of base-pairs (i.e., FP and FN) and $C(\theta)$ does not depend on the prediction y (see [12] for the proof). This means the MEA-based estimator by Do et al. possess a bias against the widely-used accuracy measures for Problem 2 (SEN, PPV and MCC of base-pairs) compared with the γ -centroid estimator. Thus, the γ -centroid estimator is theoretically superior to the MEA-based estimator by Do et al. with respect to those accuracy measures. In computational experiments, the authors confirmed that the γ -centroid estimator is always better than the MEA-based estimator when we used the same probability distribution of secondary structures. See [12] for details of the computational experiments.

A.5.3 Estimation of phylogenetic trees (Problem 4)

The γ -centroid estimator for Problem 4 can be introduced as follows:

Estimator 3 (γ -centroid estimator for Problem 4) *For Problem 4, we obtain the γ -centroid estimator (Definition 7) where the predictive space Y is equal to $\mathcal{T}(S)$ and the probability distribution on Y is taken by $p^{(t)}(\theta|S)$.*

The following property is easily obtained by Theorem 2 and [17].

Property 7 (Relation of 1-centroid estimator and topological distance) *The γ -centroid estimator with $\gamma = 1$ (i.e. centroid estimator) for Problem 4 minimizes expected topological distances.*

For $X \in I^{(0)}$ ($I^{(0)}$ is a set of partitions of S and is formally defined in the previous section), we call the marginalized probability $p_X = \sum_{\theta \in \mathcal{T}(S)} I(\theta_X = 1) p^{(t)}(\theta|S)$ *partitioning probability*. However, it is difficult to compute $\{p_X\}_{X \in I^{(0)}}$ as efficiently as in the prediction of secondary structures of RNA sequences, where it seems possible to compute the base-pairing probability matrix in polynomial time by using dynamic programming). Instead, a sampling algorithm can be used for estimating $\{p_X\}_{X \in I^{(0)}}$ approximately [16] for this problem. Once $\{p_X\}_{X \in I^{(0)}}$ is estimated, Theorem 3 leads to the following:

Property 8 (Computaion of Estimator 3) *The phylogenetic tree of Estimator 3 is found by maximizing the sum of the partitioning probabilities p_X (of the partitions given by the predicted tree) that are larger than $1/(\gamma + 1)$.*

In contrast to Estimator 1 (the γ -centroid estimator for secondary structure prediction of RNA sequence) and Estimator 2 (the γ -centroid estimator for pairwise alignment), it appears that there is no efficient method (such as dynamic programming algorithms) to computed Estimator 3 with $\gamma > 1$. Estimator 1 with $\gamma \in [0, 1]$, however, can be computed by using the following property, which is directly proven by Corollary 1 and the definition of the space $\mathcal{T}(S)$.

Property 9 (Estimator 3 with $0 < \gamma \leq 1$) *The γ -centroid estimator with $\gamma \in [0, 1]$ for Problem 4 contains its consensus estimator.*

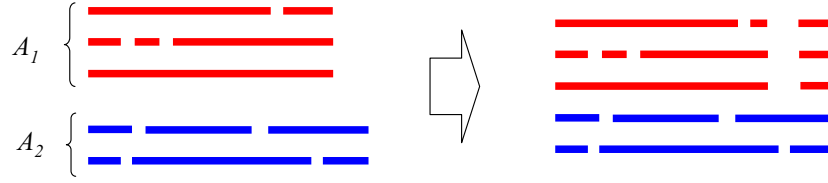


Figure S4: **Alignment between two multiple alignments A_1 and A_2 (Problem 10)**

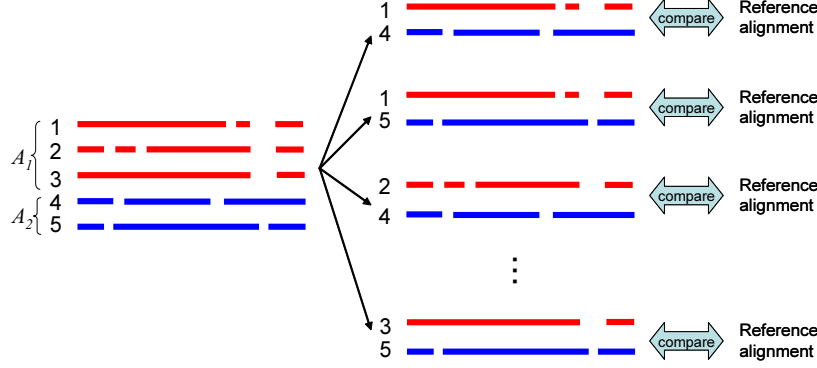


Figure S5: **An evaluation process for Problem 10.** The comparison between every pairwise alignment and the reference alignment is conducted using TP, TN, FP and FN with respect to the aligned-bases.

A.5.4 Alignment between two *alignments* of biological sequences

In this section we consider the problem of the alignment between two multiple alignments of biological sequences (Figure S4), which is often important in the multiple alignment of RNA sequences [19]. This problem is formulated as follows.

Problem 10 (Alignment between two alignments of biological sequences) *The data is represented as $D = \{A, A'\}$ where A and A' are alignments of biological sequences and the predictive space Y is equal to $\mathcal{A}(A, A')$, that is, the space of the alignments of A and A' .*

In the following, $l(A)$ and $n(A)$ denote the length of the alignment and the number of sequences in the alignment A , respectively. If both A and A' contain a single biological sequence (with no gap), Problem 10 is equivalent to conventional pairwise alignment of biological sequences (Problem 1). As in common secondary structure prediction, the representative estimator plays an important role in this application.

Estimator 4 (Representative estimator for Problem 10) *For Problem 10, we obtain the representative estimator (Definition 10). The gain function $G'(\theta^k, y)$ is the gain function of the γ -centroid estimator. The parameter space Θ is represented as a product space $\Theta = \prod_{x \in A, x' \in A'} \mathcal{A}(x, x')$ where $\mathcal{A}(x, x')$ is defined in the previous section. The probability distribution on the parameter space Θ is given by $p(\theta|D) = \prod_{x \in A, x' \in A'} p^{(a)}(\theta^{xx'}|x, x')$ for $\theta = (\theta^{xx'})_{x \in A, x' \in A'} \in \Theta$ where $p^{(a)}(\theta|x, x')$ is given in the previous section (when x or x' contains some gaps, $p^{(a)}(\theta|x, x')$ is defined by the sequences with the gaps removed).*

Corollary 2 proves the following properties of Estimator 5.

Property 10 (A Relation of Estimator 4 with accuracy measures) *Estimator 4 is consistent with the accuracy process for Problem 10 that is shown in Figure S5. We compare every pairwise alignment of $x \in A$ and $x' \in A'$ with the reference alignment. These comparisons are made using TP, TN, FP and FN with respect to the aligned-bases (e.g., using SEN, PPV and F-score).*

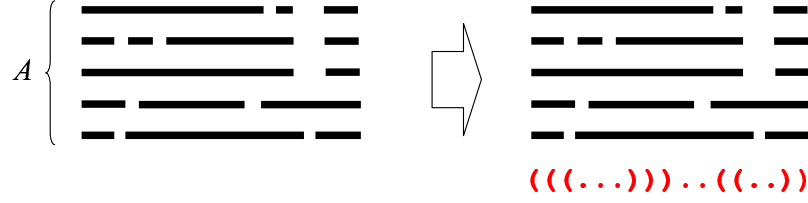


Figure S6: **Common secondary structure prediction (Problem 11)**

Property 11 (Computation of Estimator 4) *Estimator 4 can be given by maximizing the sum of probabilities \overline{p}_{ik} that are larger than $1/(\gamma + 1)$ where*

$$\overline{p}_{ik} = \frac{1}{n(A)n(A')} \sum_{x \in A} \sum_{x' \in A'} \sum_{\theta \in \Theta} I(\theta_{ik} = 1) p^{(a)}(\theta | x, x'). \quad (S10)$$

Therefore, the pairwise alignment of Estimator 4 can be computed by the Needleman-Wunsch-type DP algorithm of Eq. (S5) in which we replace p_{ij} with Eq. (S10).

Property 12 (Computation of Estimator 4 with $0 \leq \gamma \leq 1$) *The Estimator 4 with $\gamma \in [0, 1]$ contains the consensus estimator. Moreover, the consensus estimator is identical to the estimator $y = \{y_{ik}^*\}_{1 \leq i \leq l(A), 1 \leq k \leq l(A')}$:*

$$y_{ik}^* = \begin{cases} 1 & \text{if } \overline{p}_{ik} > \frac{1}{\gamma+1} \\ 0 & \text{if } \overline{p}_{ik} \leq \frac{1}{\gamma+1} \end{cases} \quad \text{for } i = 1, 2, \dots, l(A), k = 1, 2, \dots, l(A')$$

where \overline{p}_{ik} is defined in Eq. (S10).

The probability matrix $\{\overline{p}_{ik}\}_{1 \leq i \leq l(A), 1 \leq k \leq l(A')}$ is often called an *averaged aligned-base (matching) probability matrix* of A and A' . In the iterative refinement of the PROBCONS [19] algorithm, the existing multiple alignments are randomly partitioned into two groups and those two multiple alignments are re-aligned. This procedure is equivalent to Problem 10.

The estimator used in PROBCONS is identical to Estimator 4 in the limit $\gamma \rightarrow \infty$. Therefore, the estimator used in PROBCONS is a special case of Estimator 4 and it only takes into account the SEN or SPS (sum-of-pairs score) of a predicted alignment.

A.5.5 Common secondary structure prediction from a multiple alignment of RNA sequences

Common secondary structure prediction from a given multiple alignment of RNA sequences plays important role in RNA research including non-coding RNA (ncRNA) [45] and viral RNAs [46], because it is useful for phylogenetic analysis of RNAs [47] and gene finding [45, 48–50]. In contrast to conventional secondary structure prediction of RNA sequences (Problem 2), the input of common secondary structure prediction is a multiple alignment of RNA sequences and the output is a secondary structure whose length is equal to the length of the input alignment (see Figure S6).

Problem 11 (Common secondary structure prediction) *The data is represented as $D = \{A\}$ where A is a multiple alignment of RNA sequences and the predictive space Y is identical to $\mathcal{S}(A)$ (the space of secondary structures whose length is equal to the alignment).*

The representative estimator (Definition 10) directly gives an estimator for Problem 11.

Estimator 5 (The representative estimator for Problem 11) *For Problem 11, we obtain the representative estimator (Definition 10) as follows. The gain function $G'(\theta^k, y)$ is the gain function of the γ -centroid estimator. The parameter space is equal to $\Theta = \prod_{x \in A} \mathcal{S}(x)$ where $\mathcal{S}(x)$ is the space of secondary structures. The probability distribution on Θ is given by $p(\theta | D) = \prod_{x \in A} p_x(\theta^x | A)$ where $p_x(\theta^x | A)$ is the probability distribution of the secondary structures of $x \in A$ after observing the alignment A .*

For example, $p_x(\theta^x | A)$ can be given by extending the $p^{(s)}(\theta | x)$, although we have also proposed more appropriate probability distribution (see [25] for the details).

Corollary 2 proves the following properties of Estimator 5.

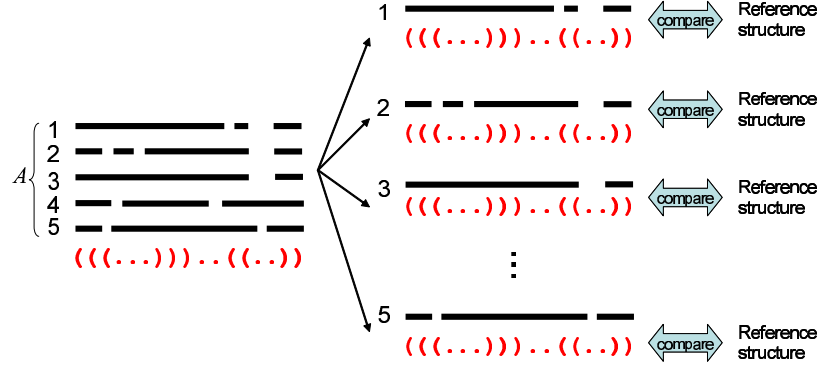


Figure S7: **An evaluation process for common secondary structure prediction (Problem 11).** The comparison between each secondary structure and the reference secondary structure is done using TP, TN, FP and FN with respect to the base-pairs.

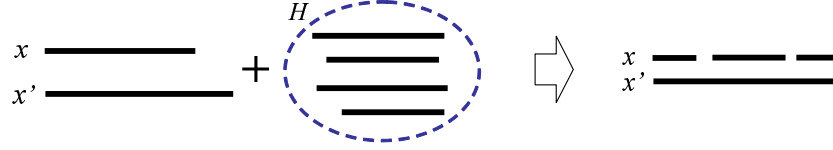


Figure S8: **Pairwise alignment using homologous sequences (Problem 12)**

Property 13 (A relation of Estimator 5 with accuracy measures) *Estimator 5 is consistent with an evaluation process for common secondary structure prediction: First, we map the predicted common secondary structure into secondary structures in the multiple alignment, and then the mapped structures are compared with the reference secondary structures based on TP, TN, FP and FN of the base-pairs using, for example, SEN, PPV and MCC (Figure S7).*

Much research into common secondary structure prediction employs the evaluation process in Figure S7 (e.g., [51]).

Property 14 (Computation of Estimator 5) *The common secondary structure of Estimator 5 is given by maximizing the sum of the averaged base-pairing probabilities \overline{p}_{ij} where*

$$\overline{p}_{ij} = \frac{1}{|A|} \sum_{x \in A} p_x(\theta_{ij}^x = 1 | A). \quad (\text{S11})$$

Therefore, the common secondary structure of the estimator can be computed using the dynamic programming algorithm in Eq. (10) if we replace p_{ij} with \overline{p}_{ij} .

Also, we can predict the secondary structure of Estimator 5 without conducting Nussinov-style DP:

Property 15 (Computation of Estimator 5 with $0 \leq \gamma \leq 1$) *The secondary structure of Estimator 5 with $\gamma \in [0, 1]$ can be predicted by collecting the base-pairs whose averaged base-pairing probabilities are larger than $1/(\gamma + 1)$.*

It should be noted that the tools of common secondary structure prediction, RNAALIFOLD [51], PETFOLD [8] and MCCASKILL-MEA [7] are also considered as a representative estimators (Definition 10). In [25], the authors systematically discuss those points. See [25] for details.

A.5.6 Pairwise alignment using homologous sequences

As in the previous application to RNA secondary structure prediction using homologous sequences, if we obtain a set of homologous sequences H for the target sequences x and x' (see Figure S8), we would have more accurate estimator for the pairwise alignment of x and x' than Estimator 1. The problem is formulated as follows.

Problem 12 (Pairwise alignment using homologous sequences) *The data is represented as $D = \{x, x', H\}$ where x and x' are two biological sequences that we would like to align, and H is a set of homologous sequences for x and x' . The predictive space Y is given by $Y = \mathcal{A}(x, x')$ which is the space of the pairwise alignments of two sequences x and x' .*

The difference between Problem 1 and this problem is that we can use other biological sequences (that seem to be homologous to x and x') besides the two sequences x and x' which are being aligned.

We can introduce the probability distribution (denoted by $p^{(a)}(\theta|x, x', h)$) on the space of multiple alignments of three sequences x , x' and h (denoted by $\mathcal{A}(x, x', h)$ and whose definition is similar to that of $\mathcal{A}(x, x')$) by a model such as the triplet HMM (which is similar to the pair HMM). Then, we obtain a probability distribution on the space of pairwise alignments of x and x' (i.e., $\mathcal{A}(x, x')$) by marginalizing $p^{(a)}(\theta|x, x', h)$ into the space $\mathcal{A}(x, x')$:

$$p(\theta|x, x') = \sum_{\theta' \in \Phi^{-1}(\theta)} p^{(a)}(\theta'|x, x', h) \quad (\text{S12})$$

where Φ is the projection from $\mathcal{A}(x, x', h)$ into $\mathcal{A}(x, x')$. Moreover, by averaging these probability distributions over $h \in H$, we obtain the following probability distribution on $\mathcal{A}(x, x')$:

$$p(\theta|x, x') = \frac{1}{|H|} \sum_{h \in H} \sum_{\theta' \in \Phi^{-1}(\theta)} p^{(a)}(\theta'|x, x', h) \quad (\text{S13})$$

where $|H|$ is the number of sequences in H .

The γ -centroid estimator with the distribution in Eq. (S13) directly gives an estimator for Problem 12. However, to compute the aligned-base-pairs (matching) probabilities p_{ik} with respect to this distribution demands a lot of computational time, so we employ the approximated γ -type estimator (Definition 12) of this γ -centroid estimator as follows.

Estimator 6 (Approximated γ -type estimator for Problem 12) *We obtain the approximated γ -type estimator (Definition 12) for Problem 12 with the following settings. The parameter space is given by $\Theta = \Theta' \times \Theta'^{\perp}$ where*

$$\Theta' = \mathcal{A}(x, x') (= Y) \text{ and } \Theta'^{\perp} = \prod_{h \in H} [\mathcal{A}(x, h) \times \mathcal{A}(x', h)]$$

and the probability distribution on the parameter space Θ' is defined by

$$p(\theta|D) = p^{(a)}(\theta^{xx'}|x, x') \prod_{h \in H} \left[p^{(a)}(\theta^{xh}|x, h) p^{(a)}(\theta^{x'h}|x', h) \right] \quad (\text{S14})$$

for $\theta = (\theta^{xx'}, \{\theta^{xh}, \theta^{x'h}\}_{h \in H}) \in \Theta = \Theta' \times \Theta'^{\perp}$. The pointwise gain function (see Definition 4) in Eq. (11) is defined by

$$\delta_{ik}(\theta) = \frac{1}{1 + |H|} \left\{ I(\theta_{ik}^{xx'} = 1) + \sum_{h \in H} \sum_{v=1}^{|h|} I(\theta_{iv}^{xh} = 1) I(\theta_{kv}^{x'h} = 1) \right\} \quad (\text{S15})$$

where $|h|$ is the length of the sequence h .

Property 16 (Computation of Estimator 6) *The alignment of Estimator 6 is equal to the alignment that maximizes the sum of p_{ik} larger than $1/(\gamma + 1)$ where*

$$p_{ik} = \frac{1}{|H| + 1} \left\{ p(\theta_{ik}^{xx'} = 1|x, x') + \sum_{h \in H} \sum_{v=1}^{|h|} p^{(a)}(\theta_{iv}^{xh} = 1|x, h) p^{(a)}(\theta_{kv}^{x'h} = 1|x', h) \right\}. \quad (\text{S16})$$

Therefore, the recursive equation of the dynamic program to calculate the alignment of Estimator 6 is given by replacing p_{ik} in Eq. (S5) with Eq. (S16).

Moreover, by using Theorem 1, we have the following proposition, which enables us to compute the proposed estimator for $\gamma \in [0, 1]$ without using (Needleman-Wunsch-type) dynamic programming.



Figure S9: RNA secondary structure prediction using homologous sequences (Problem 13)

Property 17 (Computation of Estimator 6 for $0 \leq \gamma \leq 1$) *The pairwise alignment of Estimator 6 with $\gamma \in [0, 1]$ can be predicted by collecting the aligned-bases whose probability p_{ik} in (S16) is larger than $1/(\gamma + 1)$.*

It should be noted that $\{p_{ik}\}_{1 \leq i \leq |x|, 1 \leq k \leq |x'|}$ is identical to the *probability consistency transformation* (PCT) of x and x' [19]. In PROBCONS [19], the pairwise alignment is predicted by the Estimator 6 with sufficiently large γ . Therefore, the estimator for Problem 12 used in the PROBCONS algorithm is a special case of Estimator 6.

A.5.7 RNA secondary structure prediction using homologous sequences

If we obtain a set of homologous RNA sequences for the target RNA sequence, we might have a more accurate estimator [23] for secondary structure prediction than the γ -centroid estimator (Estimator 2). This problem is formulated as follows and was considered in [23] for the first time (See Figure S9).

Problem 13 (RNA secondary structure prediction using homologous sequences) *The data D is represented as $D = \{x, H\}$ where x is the target RNA sequence for which we would like to make secondary structure predictions and H is the set of its homologous sequences. The predictive space Y is identical to $\mathcal{S}(x)$, the space of the secondary structures of an RNA sequence x .*

The difference between this problem and Problem 2 is that we are able to employ homologous sequence information for predicting the secondary structure of the target RNA sequence. In this problem, it is natural that we assume the target sequence x and each homologous sequence $h \in H$ share *common* secondary structures. The common secondary structure is naturally modeled by a *structural alignment* (that considers not only the alignment between bases but also the alignment between base-pairs), and the probability distribution (denoted by $p^{(sa)}(\theta|x, x')$) on the space of the structural alignments of two RNA sequences x and x' (denoted by $\mathcal{SA}(x, x')$) is given by the Sankoff model [52]. By marginalizing the distribution $p^{(sa)}$ into the space of secondary structures $\mathcal{S}(x)$ of the target sequence x , we obtain more reliable distribution $p(\theta|x)$ on $\mathcal{S}(x)$:

$$p(\theta|x) = \sum_{\theta' \in \Phi^{-1}(\theta)} p^{(sa)}(\theta'|x, h) \quad (\text{S17})$$

where Φ is the projection from $\mathcal{SA}(x, h)$ into $\mathcal{S}(x)$. Moreover, by averaging these probability distributions on $\mathcal{S}(x)$, we obtain the following probability distribution of secondary structures of the target sequence.

$$p(\theta|x) = \frac{1}{|H|} \sum_{h \in H} \sum_{\theta' \in \Phi^{-1}(\theta)} p^{(sa)}(\theta'|x, h) \quad (\text{S18})$$

where $|H|$ is the number of sequences in H . The γ -centroid estimator with the probability distribution in Eq. (S18) gives a reasonable estimator for Problem 13, because Eq. (S18) considers consensus secondary structures between x and $h \in H$. However, the calculation of the γ -estimator requires huge computational cost because it requires $O(nL^6)$ for computing the base-pairing probability matrix $\{p_{ik}\}$ where $p_{ik} = \sum_{\theta \in \mathcal{S}(x)} I(\theta_{ij} = 1) p(\theta|x)$ with the distribution of Eq. (S18). Therefore, we employ the approximated γ -type estimator (Definition 12) of the γ -centroid estimator, which is equivalent to the estimator proposed in [23].

Estimator 7 (Approximated γ -type estimator for Problem 13) *We obtain the approximated γ -type estimator (Definition 12) for Problem 13 with the following settings. The parameter space is given*

by $\Theta = \Theta' \times \Theta'^{\perp}$ where

$$\Theta' = \mathcal{S}(x)(= Y) \text{ and } \Theta'^{\perp} = \prod_{h \in H} [\mathcal{A}(x, h) \times \mathcal{S}(h)],$$

and the probability distribution on Θ is defined by

$$p(\theta|D) = p^{(s)}(\theta^x|x) \prod_{h \in H} [p^{(a)}(\theta^{xh}|x, h)p^{(s)}(\theta^h|h)]$$

for $\theta = (\theta^x, \{\theta^{xh}, \theta^h\}_{h \in H}) \in \Theta = \Theta' \times \Theta'^{\perp}$. Moreover, Eq. (11) in the pointwise gain function is defined by

$$\delta_{ij}(\theta) = \alpha I(\theta_{ij}^x = 1) - \frac{1 - \alpha}{|H|} \sum_{h \in H} \sum_{k < l} I(\theta_{ik}^{xh} = 1) I(\theta_{jl}^{xh} = 1) I(\theta_{kl}^h = 1)$$

for $\alpha \in [0, 1]$.

It should be noted that Estimator 13 is equivalent to the estimator proposed in [23]. The secondary structure of the estimator can be computed by the following method.

Property 18 (Computation of Estimator 7) *The secondary structure of Estimator 7 is computed by maximizing the sum of p_{ij} larger than $1/(\gamma + 1)$ where*

$$p_{ij} = \alpha p_{ij}^{(s,x)} + \frac{1 - \alpha}{|H|} \sum_{h \in H} \sum_{k < l} p_{ik,jl}^{(a,x,h)} p_{kl}^{(s,h)}. \quad (\text{S19})$$

Here, $p_{ij}^{(s,x)} = p^{(s)}(\theta_{ij}^x = 1|x)$ and $p_{ik,jl}^{(a,x,h)} = p^{(a)}(\theta_{ik}^{xh} = 1, \theta_{jl}^{xh} = 1|x, h)$. Therefore, the secondary structure of Estimator 7 can be computed by the Nussinov-type DP of Eq. (10) in which we replace p_{ij} by Eq. (S19).

The computational cost with respect to time for computing the secondary structure of Estimator 7 is $O(nL^4)$ where n is the number of RNA sequences and L is the length of RNA sequences. In [23], we employed a further approximation of the estimator, and reduced the computational cost to $O(nL^3)$. We implemented this estimator in software called CENTROIDHOMFOLD. See [23] for details of the theory and results of computational experiments. Although the authors did not mention it in their paper [23], the following property holds.

Property 19 (Computation of Estimator 7 with $0 \leq \gamma \leq 1$) *Estimator 7 with $\gamma \in [0, 1]$ can be predicted by collecting the aligned-bases where the (pseudo-)base-pairing probability of Eq. (S19) is larger than $1/(\gamma + 1)$.*

A.5.8 Pairwise alignment of *structured* RNAs

In this section, we focus on the pairwise alignment of structured RNAs. This problem is formulated as Problem 1, so the output of the problem is a usual alignment (contained in $\mathcal{A}(x, x')$). In contrast to the usual alignment problem, we can consider not only nucleotide sequences but also secondary structures in each sequence for the problem. Note that this does *not* mean the structural alignment [52] of RNA sequences, because the structural alignment produces both alignment and the common secondary structure simultaneously.

The probability distributions $p^{(a)}(\theta|x, x')$ on $\mathcal{A}(x, x')$ described in the previous section are not able to handle secondary structures of each RNA sequence. In order to obtain a probability distribution on $\mathcal{A}(x, x')$ that considers secondary structure, we employ the marginalization of the Sankoff model [52] that gives a probability distribution (denoted by $p^{(sa)}(\theta|x, x')$) on the space of possible structural alignments between two RNA sequences (denoted by $\mathcal{SA}(x, x')$). In other words, we obtain a probability distribution on the space $\mathcal{A}(x, x')$ by marginalizing the probability distribution of *structural* alignments of two RNA sequences (given by the Sankoff model) into the space $\mathcal{A}(x, x')$ as follows.

$$p(\theta|x, x') = \sum_{\theta' \in \Phi^{-1}(\theta)} p^{(sa)}(\theta'|x, x') \quad (\text{S20})$$

where Φ is the projection from $\mathcal{SA}(x, x')$ into $\mathcal{A}(x, x')$, $\theta \in \mathcal{A}(x, x')$ and $\theta' \in \mathcal{SA}(x, x')$. The difference between this marginalized probability distribution and the distributions such as Miyazawa model is that the former considers secondary structures of each sequence (more precisely, the former considers the common secondary structure).

Then, the γ -centroid estimator with this distribution Eq. (S20) will give a reasonable estimator for the pairwise alignment of two RNA sequences. However, the computation of this estimator demands huge computational cost because it uses the Sankoff model (cf. it requires $O(L^6)$ time for computing the matching probability matrix of structural alignments). Therefore, we employed the approximated γ -type estimator (Definition 12) of the γ -centroid estimator with the marginalized distribution as follows.

Estimator 8 (Approximated γ -type estimator for Problem 1 with two RNA sequences) *In Problem 1 where x and x' are RNA sequences, we obtain the approximated γ -type estimator (Estimator 2) with the following settings. The parameter space is given by $\Theta = \Theta' \times \Theta'^\perp$ where*

$$\Theta' = \mathcal{A}(x, x') (= Y), \quad \Theta'^\perp = \mathcal{S}(x) \times \mathcal{S}(x')$$

and the probability distribution on the parameter space Θ is defined by

$$p(\theta|x, x') = p^{(a)}(\theta^{(a, x, x')}|x, x')p^{(s)}(\theta^{(s, x)}|x)p^{(s)}(\theta^{(s, x')}|x')$$

for $\theta = (\theta^{(a, x, x')}, \theta^{(s, x)}, \theta^{(s, x')}) \in \Theta$. The pointwise gain function of Eq. (11) is defined by

$$\delta_{uv}(\theta) = w_1\theta_{uv}^{(a, x, x')} + w_2(\bar{R}_{uv}(\theta) + \bar{L}_{uv}(\theta')) + w_3\eta_u^{(x)}\eta_v^{(x')}$$

where

$$\begin{aligned} \bar{R}_{uv}(\theta) &:= \sum_{j: u < j, l: v < l} \theta_{uj}^{(s, x)} \theta_{vl}^{(s, x')} \theta_{jl}^{(a, x, x')}, \\ \bar{L}_{uv}(\theta) &:= \sum_{i: i < u, k: k < v} \theta_{iu}^{(s, x)} \theta_{kv}^{(s, x')} \theta_{ik}^{(a, x, x')}, \\ \eta_u^{(x)} &:= \prod_{j: u < j} (1 - \theta_{uj}^{(s, x)}) \prod_{j: j < u} (1 - \theta_{ju}^{(s, x)}), \end{aligned}$$

and w_1, w_2 and w_3 are positive weights that satisfy $w_1 + w_2 + w_3 = 1$.

This approximated γ -type estimator is equivalent to the estimator proposed in [26] and the alignment of the estimator can be computed by the following property.

Property 20 (Computation of Estimator 8) *The alignment of Estimator 8 can be computed by maximizing the sum of probabilities p_{uv} that are larger than $1/(\gamma + 1)$ where*

$$\begin{aligned} p_{uv} &= w_1 p_{uv}^{(a, x, x')} + \\ &w_2 \left(\sum_{j: u < j, l: v < l} p_{uj}^{(s, x)} p_{vl}^{(s, x')} p_{jl}^{(a, x, x')} + \sum_{i: i < u, k: k < v} p_{iu}^{(s, x)} p_{kv}^{(s, x')} p_{ik}^{(a, x, x')} \right) + w_3 q_u^{(s, x)} q_v^{(s, x')}. \end{aligned} \quad (\text{S21})$$

Here, we define

$$\begin{aligned} p_{ij}^{(s, x)} &= \sum_{\theta \in \mathcal{S}(x)} \theta_{ij} p^{(s)}(\theta|x), \\ q_u^{(s, x)} &= 1 - \sum_{i: i < u} p_{iu}^{(s, x)} - \sum_{j: u < j} p_{uj}^{(s, x)} \text{ and} \\ p_{uv}^{(a, x, x')} &= \sum_{\theta \in \mathcal{A}(x, x')} \theta_{uv} p^{(a)}(\theta|x, x'). \end{aligned}$$

Therefore, the pairwise alignment of Estimator 8 can be computed by a Needleman-Wunsch-type dynamic program of Eq. (S5) in which we replace p_{ij} with Eq. (S21).

Note that p_{uv} in Eq. (S21) is considered as a *pseudo*-aligned base probability where x_u aligns with x_v .

By checking Eq. (14), we obtain the following property:

Property 21 (Computation of Estimator 8 with $0 \leq \gamma \leq 1$) *The pairwise alignment of Estimator 8 can be predicted by collecting aligned-bases where the probability in Eq. (S21) is larger than $1/(\gamma + 1)$.*

A.6 Proofs

In this section, we give the proofs of the theorems, propositions and corollary.

A.6.1 Proof of Theorem 1

We will prove a more general case of Theorem 1 where the parameter space Θ is different from the predictive space Y and a probability distribution on Θ is assumed (cf. Assumption 2).

Theorem 4 *In Problem 3 with Assumption 1 and a pointwise gain function, suppose that a predictive space Y can be written as*

$$Y = \bigcap_{k=1}^K C_k, \quad (\text{S22})$$

where C_k is defined as

$$C_k = \left\{ y \in \{0, 1\}^n \mid \sum_{i \in I_k} y_i \leq 1 \right\} \text{ for } k = 1, 2, \dots, K$$

for an index-set $I_k \subset \{1, 2, \dots, n\}$. If the pointwise gain function in Eq. (1) (we here think θ is in a parameter space Θ which might be different from Y) satisfies the condition

$$F_i(\theta, 1) - F_i(\theta, 0) + F_j(\theta, 1) - F_j(\theta, 0) \leq 0 \quad (\text{S23})$$

for every $\theta \in \Theta$ and every $i, j \in I_k$ ($1 \leq k \leq K$), then the consensus estimator is in the predictive space Y , and hence the MEG estimator contains the consensus estimator.

(proof) It is sufficient to show that the consensus estimator $\hat{y}^{(c)}$ is contained in the predictive space Y because $\bar{G}(\hat{y}) \leq \bar{G}(\hat{y}^{(c)})$ for all \hat{y} in the MEG estimators, where

$$\bar{G}(y) := E_{\theta|D}[G(\theta, y)] = \int G(\theta, y) p(\theta|D) d\theta.$$

If we assume that $\hat{y}^{(c)}$ is not contained in the predictive space, Y that is, $\hat{y}^{(c)} \notin Y$, then there exists a k_0 such that $\hat{y}^{(c)} \notin C_{k_0}$. Because $\hat{y}^{(c)}$ is a binary vector, there exist indexes $i, j \in I_{k_0}$ such that $i \neq j$, $\hat{y}_i^{(c)} = 1$ and $\hat{y}_j^{(c)} = 1$. By the definition of $\hat{y}^{(c)}$, we obtain

$$E[F_i(\theta, 1)] > E[F_i(\theta, 0)] \text{ and } E[F_j(\theta, 1)] > E[F_j(\theta, 0)].$$

Therefore, we obtain

$$\begin{aligned} 0 &< E[F_i(\theta, 1) - F_i(\theta, 0) + F_j(\theta, 1) - F_j(\theta, 0)] \\ &= \int [F_i(\theta, 1) - F_i(\theta, 0) + F_j(\theta, 1) - F_j(\theta, 0)] p(\theta|D) d\theta \\ &\leq 0. \end{aligned}$$

In order to prove the last inequality, we use Eq. (1). This leads to a contradiction and the theorem is proved.

Remark 3 It should be noted that the above theorem holds for an arbitrary parameter space including continuous-valued spaces.

A.6.2 Proof of Theorem 2

(proof) Because $I(y_i = 1) + I(y_i = 0) = 1$ for arbitrary i , we obtain, using the definitions given in equations (S1), (S2), (S3) and (S4),

$$TP + FN = \sum_i I(\theta_i = 1) \text{ and } TN + FP = \sum_i I(\theta_i = 0).$$

Therefore, we have

$$\begin{aligned}
& \alpha_1 TP + \alpha_2 TN - \alpha_3 FP - \alpha_4 FN \\
&= (\alpha_1 + \alpha_4)TP + (\alpha_2 + \alpha_3)TN - \alpha_3 \sum_i I(\theta_i = 0) - \alpha_4 \sum_i I(\theta_i = 1) \\
&= (\alpha_2 + \alpha_3) \left(\frac{\alpha_1 + \alpha_4}{\alpha_2 + \alpha_3} TP + TN \right) - \alpha_3 \sum_i I(\theta_i = 0) - \alpha_4 \sum_i I(\theta_i = 1)
\end{aligned}$$

and this leads to the proof of the theorem.

A.6.3 Proof of Theorem 3

(proof) The expectation of the gain function of the γ -centroid estimator is computed as

$$\begin{aligned}
E_{\theta|D}[G(\theta, y)] &= \sum_{\theta \in \Theta} \sum_{i=1}^n [\gamma I(\theta_i = 1)I(y_i = 1) + I(\theta_i = 0)I(y_i = 0)] p(\theta|D) \\
&= \sum_{i=1}^n [\gamma \cdot p_i \cdot I(y_i = 1) + (1 - p_i)(1 - I(y_i = 1))] \\
&= \sum_{i=1}^n [(\gamma + 1)p_i - 1] I(y_i = 1) + \sum_i (1 - p_i)
\end{aligned}$$

where $p_i = p(\theta_i = 1|D) = \sum_{\theta \in \Theta} I(\theta_i = 1)p(\theta|D)$ is the marginalized probability. Therefore, we should always predict $y_i = 0$ whenever $p_i < 1/(\gamma + 1)$, because the assumption of Theorem 3 ensures that the prediction $y_i = 0$ never violate the condition of the predictive space Y . Theorem 3 follows by using those facts.

A.6.4 Proof of Corollary 1

(proof) For every $\theta \in \Theta$, $k = 1, 2, \dots, K$, $i, j \in J_k$, $\gamma \in [0, 1]$, we have

$$\begin{aligned}
& F_i(\theta, 1) - F_i(\theta, 0) + F_j(\theta, 1) - F_j(\theta, 0) \\
&= \gamma I(\theta_i = 1) - I(\theta_i = 0) + \gamma I(\theta_j = 1) - I(\theta_j = 0) \\
&\leq 2(I(\theta_i = 1) + I(\theta_j = 1)) - 2 \\
&\leq 0
\end{aligned}$$

and the condition of Eq. (3) in Theorem 1 is satisfied (in order to prove the last inequality, we use $I(\theta_i = 1) + I(\theta_j = 1) \leq 1$ because $i, j \in J_k$). Therefore, by Theorem 1, the γ -centroid estimator contains its consensus estimator.

The last half of the corollary is easily proved using the equation

$$\sum_{\theta \in \Theta} F_i(\theta, y_i)p(\theta|D) = \sum_{\theta \in \Theta} (I(\theta_i = y_i = 0) + \gamma I(\theta_i = y_i = 1))p(\theta|D) = \begin{cases} \gamma p_i & \text{for } y_i = 1 \\ 1 - p_i & \text{for } y_i = 0 \end{cases}$$

where $p_i = p(\theta_i = 1|D) = \sum_{\theta \in \Theta} I(\theta_i = 1)p(\theta|D)$.

A.6.5 Proof of Proposition 1

(proof) The representative estimator in Definition 10 can be written as

$$\begin{aligned}
\hat{y} &= \arg \max_{y \in Y} \int G(\theta, y)p(\theta|D)d\theta \\
&= \arg \max_{y \in Y} \int \left[\sum_{k=1}^K G'(\theta^k, y) \right] \left[\prod_{k=1}^K p^{(k)}(\theta^k|D) \right] d\theta \\
&= \arg \max_{y \in Y} \int G'(\theta', y) \left[\frac{1}{K} \sum_{k=1}^K p^{(k)}(\theta'|D) \right] d\theta'
\end{aligned}$$

Then, we finish the proof of Proposition 1.

A.6.6 Derivation of Eq. (14)

The equation is easily derived from the equality $F_i(\theta', 1) - F_i(\theta', 0) = (\gamma + 1)\delta_i(\theta') - 1$.

References

- [1] Carvalho L, Lawrence C (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc Natl Acad Sci USA* 105: 3209–3214.
- [2] Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, et al. (2009) Fast statistical alignment. *PLoS Comput Biol* 5: e1000392.
- [3] Bradley RK, Pachter L, Holmes I (2008) Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics* 24: 2677–2683.
- [4] Frith MC, Hamada M, Horton P (2010) Parameters for accurate genome alignment. *BMC Bioinformatics* 11: 80.
- [5] Do C, Woods D, Batzoglou S (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* 22: e90–98.
- [6] Lu ZJ, Gloor JW, Mathews DH (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA* 15: 1805–1813.
- [7] Kiryu H, Kin T, Asai K (2007) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics* 23: 434–441.
- [8] Seemann S, Gorodkin J, Backofen R (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res* 36: 6355–6362.
- [9] Kall L, Krogh A, Sonnhammer EL (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 21 Suppl 1: i251–257.
- [10] Kato Y, Sato K, Hamada M, Watanabe Y, Asai K, et al. (2010) RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics* 26: i460–466.
- [11] Gross S, Do C, Sirota M, Batzoglou S (2007) CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol* 8: R269.
- [12] Hamada M, Kiryu H, Sato K, Mituyama T, Asai K (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics* 25: 465–473.
- [13] Miyazawa S (1995) A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng* 8: 999–1009.
- [14] McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29: 1105–1119.
- [15] Sato K, Hamada M, Asai K, Mituyama T (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res* 37: W277–280.
- [16] Metropolis N, Rosenbluth A, Teller M, Teller E (1953) Equations of state calculations by fast computing machine. *J Chem Phys* 21: 1087–1091.
- [17] Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. *Mathematical Biosciences* 53: 131–147.
- [18] Iwasaki W, Takagi T (2010) An intuitive, informative, and most balanced representation of phylogenetic topologies. *Syst Biol* 59: 584–593.
- [19] Do C, Mahabhashyam M, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* 15: 330–340.
- [20] Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16: 412–424.
- [21] Schwartz AS, Myers EW, Pachter L (2005). Alignment metric accuracy.
- [22] Hamada M, Sato K, Asai K (2010) Prediction of RNA secondary structure by maximizing pseudo-expected accuracy. *BMC Bioinformatics* 11: 586.

- [23] Hamada M, Sato K, Kiryu H, Mituyama T, Asai K (2009) Predictions of RNA secondary structure by combining homologous sequence information. *Bioinformatics* 25: i330–338.
- [24] Ding Y, Chan C, Lawrence C (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* 11: 1157–1166.
- [25] Hamada M, Sato K, Asai K (2010) Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucleic Acids Res* : doi: 10.1093/nar/gkq792.
- [26] Hamada M, Sato K, Kiryu H, Mituyama T, Asai K (2009) CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics* 25: 3236–3243.
- [27] Roshan U, Livesay D (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* 22: 2715–2721.
- [28] Do CB, Gross SS, Batzoglu S (2006) Contralign: Discriminative training for protein sequence alignment. In: Apostolico A, Guerra C, Istrail S, Pevzner PA, Waterman MS, editors, RECOMB. Springer, volume 3909 of *Lecture Notes in Computer Science*, pp. 160–174.
- [29] Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288: 911–940.
- [30] Dowell R, Eddy S (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* 5: 71.
- [31] Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.
- [32] Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754–755.
- [33] Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis. Cambridge, UK: Cambridge University press.
- [34] Needleman S, Wunsch C (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48: 443–453.
- [35] Frith MC, Wan R, Horton P (2010) Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Res* 38: e100.
- [36] Holmes I, Durbin R (1998) Dynamic programming alignment accuracy. *J Comput Biol* 5: 493–504.
- [37] Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, et al. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 33: 121–124.
- [38] Andronescu M, Condon A, Hoos H, Mathews D, Murphy K (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics* 23: 19–28.
- [39] Nussinov R, Pieczenk G, Griggs J, Kleitman D (1978) Algorithms for loop matchings. *SIAM Journal of Applied Mathematics* 35: 68–82.
- [40] Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31: 3406–3415.
- [41] Hofacker I, Fontana W, Stadler P, Bonhoeffer S, Tacker M, et al. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh Chem* 125: 167–188.
- [42] Mathews D, Disney M, Childs J, Schroeder S, Zuker M, et al. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA* 101: 7287–7292.
- [43] Chan CY, Lawrence CE, Ding Y (2005) Structure clustering features on the Sfold Web server. *Bioinformatics* 21: 3926–3928.
- [44] Ding Y, Chan CY, Lawrence CE (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res* 32: 135–141.
- [45] Bernhart SH, Hofacker IL (2009) From consensus structure prediction to RNA gene finding. *Brief Funct Genomic Proteomic* 8: 461–471.

- [46] Schroeder SJ (2009) Advances in RNA structure prediction from sequence: new tools for generating hypotheses about viral RNA structure-function relationships. *J Virol* 83: 6326–6334.
- [47] Stocsits RR, Letsch H, Hertel J, Misof B, Stadler PF (2009) Accurate and efficient reconstruction of deep phylogenies from structured RNAs. *Nucleic Acids Res* 37: 6184–6193.
- [48] Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* 102: 2454–2459.
- [49] Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* 23: 1383–1390.
- [50] Okada Y, Sato K, Sakakibara Y (2010) Improvement of structure conservation index with centroid estimators. *Pac Symp Biocomput* : 88–97.
- [51] Bernhart S, Hofacker I, Will S, Gruber A, Stadler P (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 9: 474.
- [52] Sankoff D (1985) Simultaneous solution of the RNA folding alignment and protosequence problems. *SIAM J Appl Math* : 810–825.