Published online 15 September 2010

Nucleic Acids Research, 2011, Vol. 39, No. 2 ***393–402***
doi:10.1093/nar/gkq792

Improving the accuracy of predicting secondary structure for aligned RNA sequences

Michiaki Hamada[1,2,*], Kengo Sato[3] and Kiyoshi Asai[2,3]

[1]Mizuho Information & Research Institute, Inc, 2-3 Kanda-Nishikicho, Chiyoda-ku, Tokyo 101-8443, [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koto-ku, Tokyo 135-0064 and [3]Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan

Received May 23, 2010; Revised August 16, 2010; Accepted August 20, 2010

ABSTRACT

Considerable attention has been focused on predicting the secondary structure for aligned RNA sequences since it is useful not only for improving the limiting accuracy of conventional secondary structure prediction but also for finding non-coding RNAs in genomic sequences. Although there exist many algorithms of predicting secondary structure for aligned RNA sequences, further improvement of the accuracy is still awaited. In this article, toward improving the accuracy, a theoretical classification of state-of-the-art algorithms of predicting secondary structure for aligned RNA sequences is presented. The classification is based on the viewpoint of maximum expected accuracy (MEA), which has been successfully applied in various problems in bioinformatics. The classification reveals several disadvantages of the current algorithms but we propose an improvement of a previously introduced algorithm (CentroidAlifold). Finally, computational experiments strongly support the theoretical classification and indicate that the improved CentroidAlifold substantially outperforms other algorithms.

INTRODUCTION

Prediction of the secondary structure for aligned RNA sequences (which is usually called a 'common' or 'consensus' secondary structure) is an important problem in many fields of RNA research, including non-coding RNA (1) and viral RNAs (2). The (common) secondary structure is often useful for improving the limiting accuracy of conventional secondary structure prediction [e.g. RNAfold (3) and Mfold (4)]. Moreover, it plays an essential role in phylogenetic analysis of RNAs and gene finding of RNAs from genomic sequences (5–11).

A number of algorithms for common secondary structure prediction have been proposed. A well-known program, RNAalifold (12,13), is based on the free energies of the secondary structures of the RNA sequences in the given alignment (thermodynamic information) and the mutation of two bases that maintain a base pair (bonus of co-variation). RNAalifold has been used in a number of studies e.g. (5,6,9). Recent changes to RNAalifold have improved its performance substantially (12). A probabilistic version of RNAalifold is called RNAalipffold model (12), which gives a probability distribution of common secondary structures of the input alignment. RNAalifold is considered as the maximum likelihood (ML) estimation of RNAalipffold model. Another popular program, Pfold (14), uses stochastic context-free grammars (SCFGs) with the phylogenetic information of the input RNA sequences. Pfold also provides a probability distribution of common secondary structures of the input alignment (we call it Pfold model). Recently, PETfold (15), which employs both the thermodynamic and phylogenetic information, has been proposed. McCaskill-MEA (16) achieved robust prediction of common secondary structure by using the averaged base pairing probability matrix based on a thermodynamic model. Both PETfold and McCaskill-MEA are based on the principle of the maximum expected accuracy (MEA), which maximizes the expected accuracy of a prediction with respect to a probability distribution on the entire set of candidate solutions. Another MEA-based algorithm, CentroidAlifold (17), maximizes the sum of the expected gain (of a carefully designed gain function) under a probability distribution of secondary structures of every RNA sequence in the alignment, where the distribution is given by McCaskill model (18) (energy-based model) or CONTRAfold model (19) (machine learning-based model). The estimator of

*To whom correspondence should be addressed. Tel.: +81 3 5281 5271; Fax: +81 3 5281 5331; Email: hamada-michiaki@aist.go.jp

CentroidAlifold is closely related to the γ-centroid estimator (17), which is employed in CentroidFold (20) for conventional secondary structure prediction. The combination of RNAalipffold model (or Pfold model) with the γ-centroid is called RNAalipffold-Centroid (or Pfold-Centroid) (17).

Recent studies have suggested that the principle of MEA, which is used in PETfold, McCaskill-MEA and CentroidAlifold, gives powerful estimators for estimation problems in bioinformatics, including RNA secondary structure prediction (17,19,21,22), common secondary structure prediction from a multiple alignment of RNA sequences (15,16), pairwise/multiple alignment of biological sequences (23,24,25,26), genome alignment (27), transmembrane topology and signal peptide prediction (28), recombination detection (29), gene prediction (30), RNA–RNA interaction (31) and multiple alignment for structured RNAs (32).

In this study, toward improving the accuracy of previously proposed algorithms, we first classify existing software of common secondary structure prediction. The classification is based on an MEA-based estimator with respect to the evaluation process of the common secondary structure prediction. We then propose an improvement of CentroidAlifold by using a mixture distribution of a probability distribution of common secondary structures (e.g. RNAalifold model or Pfold model) and that of secondary structures of each RNA sequence (e.g. McCaskill model or CONTRAfold model). Finally, we show that the improved CentroidAlifold substantially outperforms other algorithms by performing computational experiments.

MATERIALS AND METHODS

Two evaluation processes for common secondary structure prediction

In the problem of common secondary structure prediction (we do not predict each secondary structure of the sequences in a given alignment but predict one common secondary structure of the alignment), two evaluation processes have been used. The first one is to compare the predicted common secondary structure with a reference (correct) common secondary structure directly (Evaluation Process 1; Supplementary Figure S1). However, this evaluation is not so often used in actual evaluations because the definition of the reference common secondary structure is unclear and it is often difficult to prepare the reference common secondary structure of a given alignment, for example, the alignment produced by aligners such as ClustalW (33) and ProbCons (34). Therefore, another evaluation is often conducted (Evaluation Process 2; Supplementary Figure S2): the (predicted) common secondary structure is mapped to each RNA sequence in the input alignment, and then the mapped secondary structures are compared with the reference secondary structure of each RNA sequence (the reference secondary structure is, e.g. obtained from X-ray crystallography or NMR). In other words, a (common) secondary structure that recovers the secondary structures of each RNA sequence in the alignment is a good prediction. Although we need to prepare the reference secondary structures of each RNA sequence in the input alignment in order to conduct this evaluation, it is much easier than preparing the reference 'common' secondary structure, as the reference structure of individual RNA sequence can be obtained by using a database, such as Rfam (35) or RNAstrand (36). It should be noted that, in the both evaluation processes, the comparison between two secondary structures (or common secondary structures) is based on the base pairs that are essential for forming secondary/tertiary structures, which are known to be biologically important. More precisely, the sensitivity (SEN) and positive predictive value (PPV) with respect to base-pairs are commonly used in those evaluations.

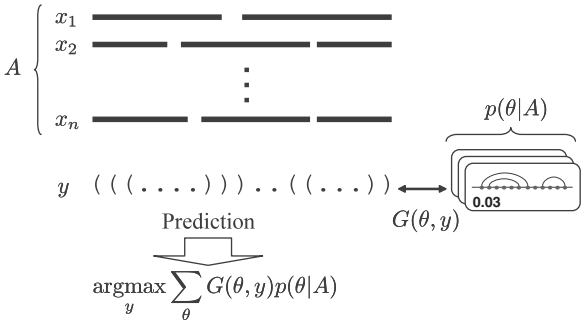

Figure 1. The MEA-based estimator (E1) with respect to Evaluation Process 1. We assume there exists a probability distribution $p(\theta|A)$ of the common secondary structures of the alignment A, and a gain function $G(\theta, y)$ between two secondary structure whose length is equal to the length of the alignment (y and θ are considered as the predicted structure and the reference structure, respectively). The gain function characterizes a similarity between the two secondary structures. The estimator is consistent with Evaluation Process 1 (Supplementary Figure S1). See Supplementary Section A.4.1 for details.

MEA-based estimators

As proposed in (17), two MEA-based estimators (of secondary structure prediction for aligned RNA sequences) that fit with the two evaluation processes can be introduced:

- **(E1)** the estimator that fits with Evaluation Process 1, which maximizes the expected gain (of a gain function) under a probability distribution of (common) secondary structures of the input alignment (Figure 1);
- **(E2)** the estimator that fits with Evaluation Process 2, which maximizes the 'sum' of the expected gain (of a gain function) under a probability distribution of secondary structures of each RNA sequence in the alignment (Figure 2).

In the above estimators, the 'gain function' characterizes a similarity between a predicted structure and the reference structure, and should fit with the accuracy measures for the target problems. Also, the probability distributions play an important role in the estimator. Further details of the estimators are shown in Supplementary Section A.4.

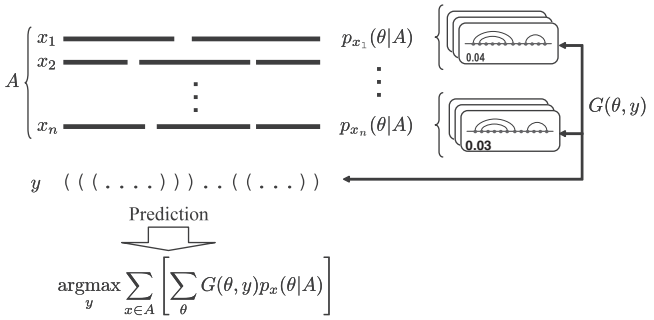

Figure 2. The MEA-based estimator (E2) with respect to Evaluation Process 2. We assume there exists a probability distribution $p_x(\theta|A)$ of common secondary structures of x for every $x \in A$ and a gain function $G(\theta, y)$ between two secondary structure whose length is equal to the length of the alignment (y and θ are considered as the predicted structure and the reference structure, respectively). The estimator is consistent with Evaluation Process 2 (see Supplementary Figure S2). See Section A.4.2 in the supplementary information for details.

Experimental settings

We used a Linux machine with a 2.8 GHz AMD Opteron Processor 854 and 64 GB of memory.

Comparison of methods. In the experiments, we compared the following algorithms or tools: (i) CentroidAlifold (new) (this work), (ii) CentroidAlifold (old) (12), (iii) RNAalifold (updated version of ViennaRNA package 1.8.3) (12), (iv) RNAalifold-Centroid (with the updated version of RNAalifold) (12,17), (v) Pfoldcentroid (14,17) and (vi) PETfold (15). In CentroidAlifold (new/old), we used two probability distributions of secondary structures of a given RNA sequence [i.e. $p(\theta|x)$ in Equation (2)]: the McCaskill model in ViennaRNA package 1.8.3 (3) and the CONTRAfold model (Version 2.02) (19). In CentroidAlifold (new), we employed two probability distributions for $p(\theta|A)$ in Equation (2): the Pfold model (14) and the RNAalipfold model (12). The weighting was fixed at $w = 1/2$. We used 17 γ parameters: $\gamma \in \{2^k : -5 \leq k \leq 10, k \in \mathbb{Z}\} \cup \{6\}$ for CentroidAlifold, Pfold-Centroid and RNAalifold-Centroid in order to draw the performance (SEN–PPV) curves.

Data sets. We used the data set of Kiryu *et al*. (16) that contains 85 multiple alignments and their reference common secondary structures. The number of families in the data set is 17; For each family, there are 5 sub alignments of randomly selected 10 sequences [This data set is the same as in our previous study (17)]. Each item in the data set consists of a manually curated multiple alignment and the reference common secondary structure of the alignment, which were derived from the Rfam database (35,37) and reliable publications. (In other words, the data set does not contain any 'predictions' at all.) The reference common secondary structure is used when we conduct Evaluation Process 1. Furthermore, the reference secondary structures of each sequence in the alignment are obtained by mapping the reference common secondary structure to the sequence. These reference structures are used for Evaluation Process 2. We produced several multiple alignments from the same sequences in the reference alignments by using four multiple aligners: ProbCons (34), MAFFT (38), MXSCARNA (39) and ClustalW (33). Those multiple alignments were used in Evaluation Process 2.

RESULTS AND DISCUSSION

Theoretical classification of state-of-the-art algorithms reveals disadvantages of those algorithms

CentroidAlifold (17), RNAalifold (12), Pfold (14), PETfold (15), RNAalifold-Centroid (12,17), Pfold-Centroid (14,17) and McCaskill-MEA (16) can be written as the MEA-based estimator (E2) in Figure 2 with a combination of the gain function G and a probability distribution $p_x(\theta|A)$ of secondary structures of x in the input alignment, as follows. [See Table 1; See also Supplementary Section A.5 for more details; Note that the estimator (E1) can be considered as the estimator (E2) as described in Supplementary Section A.4.3.]

First, the gain function G is one of the following types:

- **(G1)** the delta function (denoted by $G^{(\delta)}$) that returns 1 only when two secondary structures are 'exactly' the same;
- **(G2)** the gain function used in CONTRAfold (19) (denoted by $G_\gamma^{(\mathrm{contra})}$; See also Supplementary Equation (S10)), which is a sum of the correctly predicted (loop or base pairs) position in the sequence; and
- **(G3)** the gain function used in the γ-centroid estimator (17) (denoted by $G_\gamma^{(\mathrm{centroid})}$; See also Supplementary Equation (S6)), which is a weighted sum of the number of true-positive base pairs and true-negative base pairs

Second, the probability distribution $p_x(\theta|A)$ is one of the following types:

- **(P1)** a probability distribution of common secondary structures for the input alignment A, RNA alipffold model (12) or Pfold model (14);
- **(P2)** a probability distribution of secondary structures for individual RNA sequence x, McCaskill model (18) or CONTRAfold model (19); and
- **(P3)** a mixture of probability distributions of (P1) and (P2).

We emphasize that 'every' algorithm in Table 1 [expect for 'CentroidAlifold (new)', the proposed algorithm in this study] has drawbacks in the gain function and/or the probability distribution because there are several disadvantages of the gain function (G1) and (G2), and the probability distribution (P1) and (P2) as follows.

The use of the gain function (G1) means that the estimator is closely related to the ML estimator, and a number of recent studies have indicated that the ML estimator does not necessarily give reliable predictions for estimation problems on a high-dimensional discrete

Table 1. All cases are represented in terms of a gain function ($G(\theta, y)$) and a probability distribution of sequence x in the input alignment A ($p_x(\theta|A)$) which are components in the estimator (E2) (Figure 2)

Algorithms	$G(\theta, y)$		$p_x(\theta\|A)$		Ref.
CentroidAlifold (new)	G3	$G_\gamma^{(\mathrm{centroid})}$	P3	Mixture of $p^{(\mathrm{mcc})}(\theta\|x)/p^{(\mathrm{contra})}(\theta\|x)$ and $p^{(\mathrm{alipffold})}(\theta\|A)/p^{(\mathrm{pfold})}(\theta\|A)$	this work
CentroidAlifold (old)	G3	$G_\gamma^{(\mathrm{centroid})}$	P2	$p^{(\mathrm{mcc})}(\theta\|x)$ or $p^{(\mathrm{contra})}(\theta\|x)$	(16)
McCaskill-MEA	G2	$G_\gamma^{(\mathrm{contra})}$	P2	$p^{(\mathrm{mcc})}(\theta\|x)$	(26)
PETfold	G2	$G_\gamma^{(\mathrm{contra})}$	P3	Mixture of $p^{(\mathrm{mcc})}(\theta\|x)$ and $p^{(\mathrm{pfold})}(\theta\|A)$	(37)
Pfold	G2	$G_1^{(\mathrm{contra})}$	P1	$p^{(\mathrm{pfold})}(\theta\|A)$	(27)
Pfold-Centroid	G3	$G_\gamma^{(\mathrm{centroid})}$	P1	$p^{(\mathrm{pfold})}(\theta\|A)$	(16,27)
RNAalifold	G1	$G^{(\delta)}$	P1	$p^{(\mathrm{alipffold})}(\theta\|A)$	(2)
RNAalipffold-Centroid	G3	$G_\gamma^{(\mathrm{centroid})}$	P1	$p^{(\mathrm{alipffold})}(\theta\|A)$	(2,16)

Algorithms	Disadvantages	S.I.
CentroidAlifold (old)	No use of the information of the input alignment A	Section A.5.1
McCaskill-MEA	Use of $G_\gamma^{(\mathrm{contra})}$; no use of the information of the input alignment A	Section A.5.7
PETfold	Use of $G_\gamma^{(\mathrm{contra})}$	Section A.5.4
Pfold	Use of $G_\gamma^{(\mathrm{contra})}$	Section A.5.5
Pfold-Centroid	Use of the same distribution $p_x(\theta\|A)$ for all $x \in A$	Section A.5.6
RNAalifold	Use of G^{δ} (i.e. use of the ML estimator)	Section A.5.2
RNAalipffold-Centroid	Use the same distribution $p_x(\theta\|A)$ for all $x \in A$	Section A.5.3

$G_\gamma^{(\mathrm{centroid})}$, $G^{(\delta)}$ and $G_\gamma^{(\mathrm{contra})}$ are the gain function used in the γ-centroid estimator (17), the delta function and the gain function used in CONTRAfold (19), respectively. $p^{(\mathrm{mcc})}(\theta|x)$ and $p^{(\mathrm{contra})}(\theta|x)$ are McCaskill model (18) and CONTRAfold model (19), respectively, each of which is a probability distribution of secondary structures of RNA sequence x. $p^{(\mathrm{alipffold})}(\theta|A)$ and $p^{(\mathrm{pfold})}(\theta|A)$ are RNAalipffold model (12) and Pfold model (14), respectively, each of which is a probability distribution of common secondary structures of the alignment A. G1-3 and P1-3 show the types of the gain function and the probability distribution, respectively, and G1, G2, P1 and P2 have drawbacks (see the main text). The disadvantages of each algorithm are shown in the bottom table. For comparison, the improved CentroidAlifold [denoted by 'CentroidAlifold (new)'], which is introduced in this work, is also shown. The column 'S.I.' shows the section in the Supplementary Data.

space, because there are huge number of suboptimal solutions (40) and it is not optimized for the accuracy measures of the target problem (17). The gain function (G2) has a 'bias' to the commonly used accuracy measures of secondary structure prediction, compared to the gain function (G3) (17). More precisely, in (17), Hamada *et al*. proved

$$G_\gamma^{(\mathrm{contra})}(\theta, y) = G_\gamma^{(\mathrm{centroid})}(\theta, y) + A(\theta, y) + C(\theta) \quad (1)$$

where $A(\theta, y)$ is positive for 'false' predictions of base pairs (i.e. false positive and false negative) and $C(\theta)$ does not depend on the prediction y. This means that the gain function $G_\gamma^{(\mathrm{contra})}(\theta, y)$ has a bias against accurate predictions of base pairs in the secondary structure compared with $G_\gamma^{(\mathrm{centroid})}(\theta, y)$, so the estimator with $G_\gamma^{(\mathrm{centroid})}$ is theoretically superior to the estimator with $G_\gamma^{(\mathrm{contra})}$. See (17) for more detailed descriptions.

The use of the probability distribution (P1) has a disadvantage because the probability distribution is the 'same' for each RNA sequence x in the alignment, although it is natural that $p_x(\theta|A) \neq p_{x'}(\theta|A)$ for $x \neq x'$. A drawback of the probability distribution (P2) is that the probability distribution does not employ the information of the input multiple alignment at all. [For example, the probability distribution (P2) does not consider either the covariance bonus or the phylogenetic information of the input alignment.]

These investigations drive us to improve the current CentroidAlifold in the following.

An improvement of CentroidAlifold: theoretically better choice for gain function and probability distributions in MEA-based estimator

In order to overcome the drawbacks of the current state-of-the-art algorithms in Table 1, we employ the MEA-based estimator (E2) with the combination of the gain function (G3) and the probability distribution (P3). Note that there is no algorithm that uses this combination. This means that we replace the probability distribution in CentroidAlifold by a 'mixture' of the probability distribution of common secondary structures of A (from the RNAalipffold or Pfold model) and the probability distribution of secondary structures of the individual sequence x in A (from the McCaskill or CONTRAfold model):

$$p_x(\theta|A) = w \cdot p(\theta|x) + (1 - w) \cdot p(\theta|A) \quad (2)$$

where $w \in [0, 1]$ is a weight parameter and $p(\theta|x)$ is identical to McCaskill model or CONTRAfold model, and $p(\theta|A)$ is identical to RNAalipffold model or Pfold model. Using Equation (2), therefore, means that we consider not only the probability distribution of secondary structures of an individual RNA sequence in A but also the probability distribution of (common) secondary structures of the alignment A. If $w = 1$, the estimator is equal to that of CentroidAlifold (17) (see also Supplementary Section A.5.1.). On the other hand, if $w = 0$, the estimator is equivalent to that of RNAalipffold-Centroid or Pfold-Centroid, which are the γ-centroid estimator (17)

with the RNAalipffold or Pfold models, respectively (see also Supplementary Section A.5.3 and A.5.6).

The combination of the gain function (G3) and probability distribution (P3) is theoretically better choice than the other combinations, and this will also be confirmed by computational experiments in the following sections.

It should be noted that we used the 'same' gain function in both the previous and new CentroidAlifold because the gain function is still thought to be better than any other gain functions [including (G1) and (G2)] for predicting accurate base pairs in (common) secondary structures.

Computation of the improved CentroidAlifold

In the computation of CentroidAlifold, we first compute $(n+1)$ base pairing probability matrices, where n is the number of sequences in A: $\{p_{ij}^{(x)}\}_{i<j}$ for $x \in A$ and $\{p_{ij}^{(A)}\}_{i<j}$ where

$$p_{ij}^{(x)} = \sum_{\theta \in S} I(\theta_{ij} = 1) p(\theta|x) \text{ and}$$

$$p_{ij}^{(A)} = \sum_{\theta \in S} I(\theta_{ij} = 1) p(\theta|A).$$

[Here, $\mathcal{S} := \mathcal{S}(A) = \mathcal{S}(x)$.] The computational time for this is equal to $O(n|A|^3)$ because each base pairing probability matrix can be computed by the Inside-Outside algorithm (41). Finally, CentroidAlifold conducts the following Nussinov-style dynamic programming (DP) recursion (42).

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

where

$$p_{ij}^{*} = \frac{w}{n} \sum_{x \in A} p_{ij}^{(x)} + (1 - w) p_{ij}^{(A)} \quad (4)$$

and $M_{i,j}$ is the optimal score of the subsequence $x_{i \cdots j}$. Note that p_{ij}^{*} is derived from the mixture distribution of Equation (2). This DP algorithm means that CentroidAlifold maximizes the sum of (base pairing) probabilities p_{ij}^{*} of Equation (4) which are larger than $1/(\gamma+1)$. This DP algorithm requires $O(|A|^3)$ time and the total computational time of CentroidAlifold still remains $O(n|A|^3)$.

Implementation

The improved CentroidAlifold is in the 'same' package as CentroidFold (20). The software can employ a mixed distribution given by an arbitrary combination of RNAalipffold, Pfold, McCaskill and CONTRAfold models. The default probability distribution used in CentroidAlifold is an equally weighted mixture of the RNAalipffold and McCaskill model [$w = 1/2$ in Equation (2)].

Improved CentroidAlifold substantially outperforms other methods in computational experiments

CentroidAlifold (new) clearly outperformed the other algorithms with respect to Evaluation Process 1 for the 'reference' alignment (Figure 3). Note that we cannot apply Evaluation Process 1 to the predicted common secondary structure from the multiple alignments produced by the aligners such as ClustalW. This implies that the averaged probability distribution of $p_x(\theta|A)$ (i.e. a probability distribution of secondary structures of x) for $x \in A$ of CentroidAlifold gives a reliable probability distribution of common secondary structures of A, because, by using the averaged distribution, CentroidAlifold [the estimator (E2)] is considered as the estimator (E1) that is suitable to Evaluation Process 1. (See also Supplementary Section A.4.3.)

Moreover, CentroidAlifold (new) outperformed the other algorithms with respect to Evaluation Process 2 (Figure 4 and Supplementary Figures S1–S3). Precisely speaking, CentroidAlifold (new) (the solid lines in red and green colors in each figure) clearly improved the performances of CentroidAlifold (old) (the dashed lines in red and green colors in each figure), and both RNAalipffold-Centroid and Pfold-Centroid (the blue lines), which indicated that the mixing distribution used in CentroidAlifold (new) works very well. Especially, the maximum sensitivity of CentroidAlifold is much better than the one of the other algorithms (Table 2).

In CentroidAlifold (new), there is few difference of performances between the use of McCaskill model and CONTRAfold model, while the former is about two times as fast as the latter (Table 3). This is because the time for computing the base pairing probability matrix with the CONTRAfold model is longer than for the McCaskill model implemented in the ViennaRNA package. Table 3 also indicates that RNAalifold is the fastest tool because the computational cost of RNAalifold, unlike that of the other tools, does not depend on the number of RNA sequences in the alignments. Moreover, RNAalifold need not use the Inside-Outside algorithm for computing the base pairing probability matrix but can use a Nussinov-type algorithm (42) [cf. Equation (3)] for computing the consistent (common) secondary structure, while CentroidAlifold, PETfold, RNAalipffold-Centroid and Pfold must use both (see 'Materials and Methods' section for details). As a result, RNAalifold is more than 10 times faster than the other software and algorithms (Table 3).

By the results of our benchmark (Supplementary Table S1), it seems to be enough to use $\gamma = 2$ or 4 for obtaining the common secondary structure that achieves a balance between SEN and PPV [i.e. that has a favorable Matthews Correlation Coefficient (MCC)]. Moreover, we tried common secondary structure prediction by combining CentroidAlifold with the 'pseudo'-expected MCC (M. Hamada, K. Sato and K. Asai, submitted for publication) In the manuscript, we found that the pseudo-expected MCC of a given secondary structure can be computed efficiently (while there is no efficient method to compute the expected MCC) and that the pseudo-expected

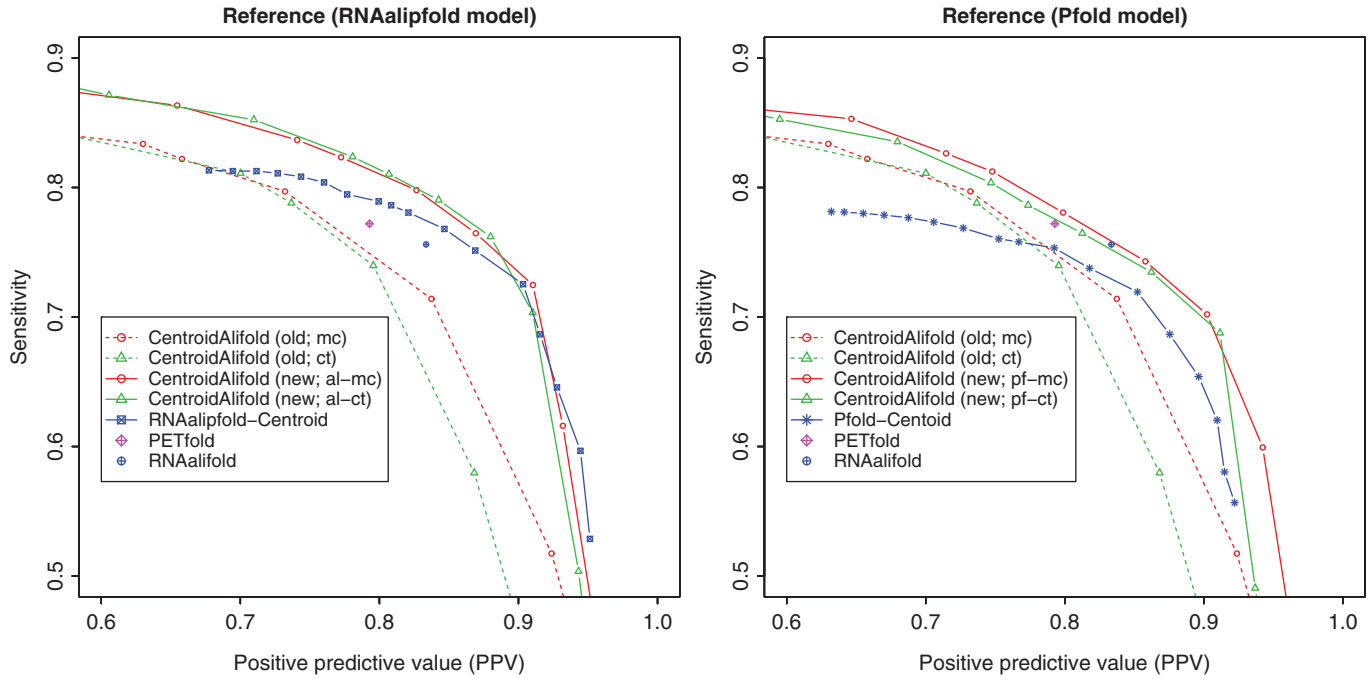

Figure 3. The performance of common secondary structure prediction with the reference alignments with respect to Evaluation Process 1. The horizontal and vertical axes indicate PPV and SEN, respectively. Better performances are in the upper-right areas of each figure (worse performances are to the lower left). The results for the RNAalipffold model are shown on the left and those for the Pfold model on the right. The labels 'mc', 'ct', 'pf' and 'al' indicate the McCaskill, CONTRAfold, Pfold and RNAalipffold models, respectively. CentroidAlifold (old: X) indicates CentroidAlifold with probability distribution X (where X = 'mc' or 'ct'). CentroidAlifold (new: Y-X) indicates CentroidAlifold with a mixture of the probability distributions X and Y where Y is $p(\theta|A)$, X is $p(\theta|x)$ and $w = 1/2$ in Equation (2) (Y = 'pf' or 'al'). The dashed lines (red/green) show the performance curves of the previous CentroidAlifold, while the solid lines (red/green) show the performance curves of the new CentroidAlifold. In both figures, the performances of PETfold and RNAalifold are also shown.

MCC is a reliable approximation to the expected MCC. By using the pseudo-expected MCC, we are able to select the secondary structure from among the several secondary structures based on 17 values of γ that are used for drawing the performance curves of CentroidAlifold. (Note that we did *not* use the correct structures for the selection.) Table 4 indicates that the selected common secondary structures achieved better MCC than the structures predicted by PETfold and RNAalifold. On the other hand, Table 3 shows that there was only small computational overhead using the pseudo-expected MCC, compared with the prediction with a fixed γ. This is due to the fact that computing the base pairing probability matrices is the dominant factor in the computational time of CentroidAlifold.

The computational experiments indicated that a good choice (with respect to accuracy and speed) of the probability distribution $p_x(\theta|A)$ in the estimator (E2) is a mixture of the RNAalipffold and the McCaskill model. Also, we have performed the experiments using various parameters of weight [i.e. w in Equation (2)] and confirmed that a weight parameter of around 0.5 in the mixture distribution generally gives a good performance (Figure 5 and Supplementary Figures S6–S10). There is, however, room to do research on the probability distribution ($p_x(\theta|A)$ in the estimator (E2) [cf. Equation (S3) in Supplementary Data], because there are a number of possibilities to obtain a better mixture of distributions than the one used in the new CentroidAlifold. For example, we can employ a mixture of three probability distributions [the McCaskill model (18), the Pfold (14) model and the RNAalipffold (12) model], thereby considering the thermodynamic information, the phylogenetic information and the covariance bonus. We can also try the McCaskill model implemented in the software RNAstructure (22) that employs more elaborate energy models than the Vienna RNA package. Finding better combinations of the probability distributions in (P3) is an interesting task.

Comparison of performances among gain functions

The performance of RNAalipffold-Centroid (blue lines in the top figures in Figure 4, Supplementary Figures S1–S3) is slightly better than the performance of RNAalifold (blue points in the top figures in Figure 4, Supplementary Figures S1–S3). This shows that the γ-centroid estimator with the RNAalipffold model (that considers probability distributions of all the secondary structures) is better than the ML estimator with the RNAalipffold model, that is, RNAalifold (that only considers the optimal solution with the highest probability). This result is consistent with the theoretical results: the use of gain function (G3) is better than that of (G1).

On the other hand, CentroidAlifold with the Pfold model and the McCaskill model ['CentroidAlifold (new; pf-mc)'] is nearly equivalent to PETfold. More precisely, if we substitute the gain function (G3) into (G2) in

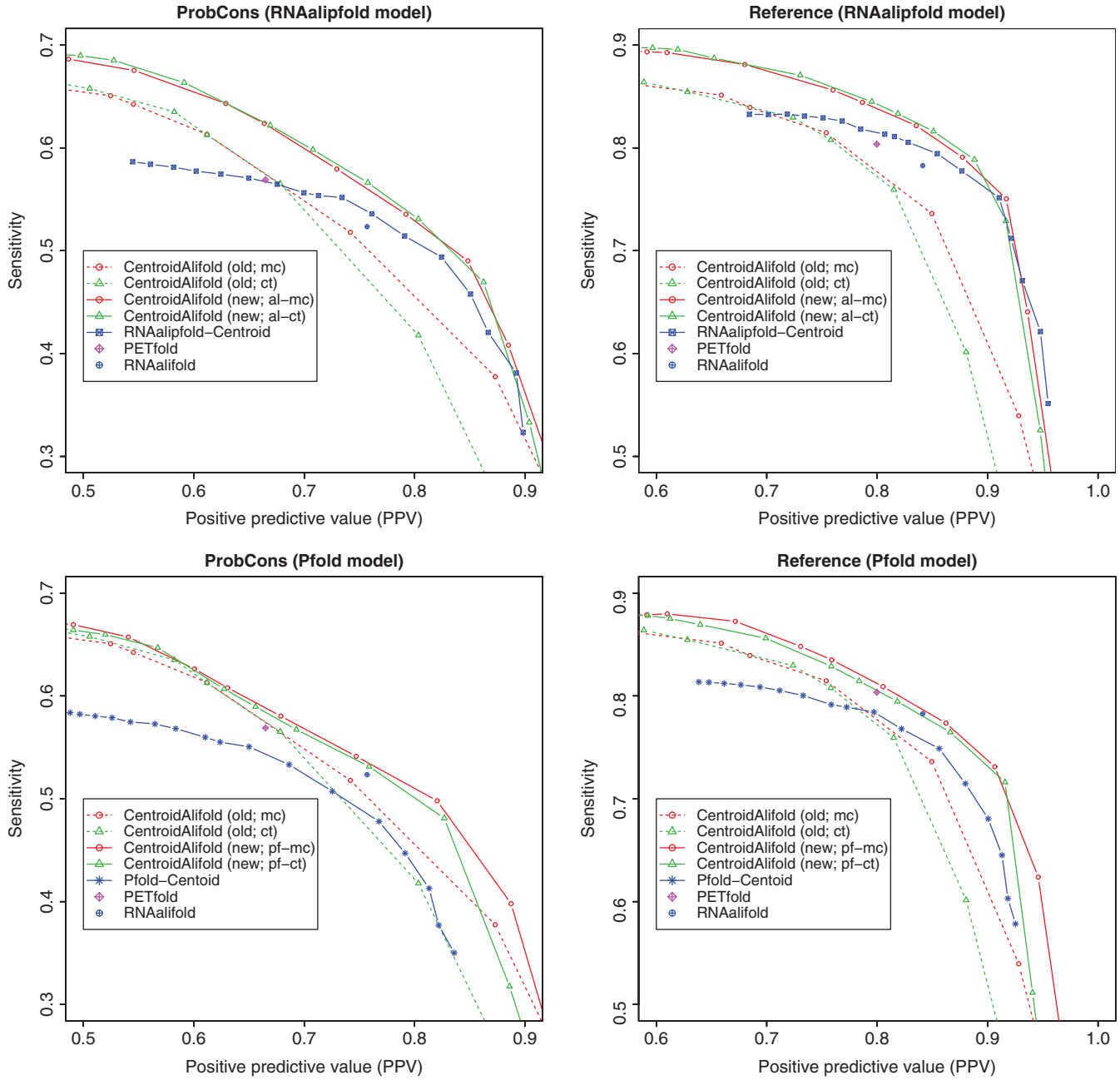

Figure 4. The performance of common secondary structure prediction for Evaluation Process 2 with alignments produced by ProbCons (left column) and the reference alignments (right column). In CentroidAlifold, we used the RNAalipffold model (top row) and the Pfold model (bottom row). See the caption of Figure 3 for notation. Also see Supplementary Figures S1–S3 for the performance with alignments produced by ClustalW (33), MAFFT (38) and MXSCARNA (39), respectively.

CentroidAlifold, the estimator is equivalent to the (main part of) PETfold. In Figure 4, it can be seen that 'CentroidAlifold (new; pf-mc)' outperforms PETfold. This confirms our theoretical results: the gain function (G2) used in PETfold contains a bias against SEN and PPV, compared with the gain function (G3) used in CentroidAlifold.

Importance of credibility (confidence) limit

Although the proposed estimator employs the entire distributions of (common) secondary structures, it still gives a 'point' estimation in a high-dimensional discrete space, and the prediction is 'uncertain' (43,44). Therefore, a global measure of uncertainty is important. Fortunately, there exist several studies related to this: the credibility (confidence) limit (43,44), which is the minimum Hamming distance of a hyper-sphere containing a specified fraction of the Boltzmann weighted ensemble. The credibility limit of a common secondary structure predicted by CentroidAlifold can be estimated using a stochastic sampling from the Boltzmann weighted ensemble of the common secondary structure. The stochastic sampling is conducted by a similar method

Table 2. The maximum sensitivity for CentroidAlifold (we used a mixture of the probability distributions of the RNAalipffold model and the McCaskill model with the same weight), RNAalipffold-Centroid and Pfold-Centroid in the SEN–PPV curves

Alignment	CentroidAlifold	RNAalipffold-Centroid	Pfold-Centroid
Reference	**0.90**	0.83	0.81
ClustalW	**0.58**	0.45	0.44
ProbCons	**0.69**	0.59	0.58
MAFFT	**0.72**	0.64	0.64
MXSCARNA	**0.75**	0.68	0.67

Evaluation Process 2 was used in this experiments. The bold values indicate the best values among three algorithms.

Table 3. Total computational time in seconds

			CentroidAlifold		
	$p(\theta\|A)$	$p(\theta\|x)$	time (all)	time (MCC)	time (fixed)
1	pf	ct	1666	1669	1626
2	pf	mc	1239	1247	1200
3	al	ct	932	929	893
4	al	mc	506	500	467
5	pf	–	869	867	832
6	al	–	133	134	99
7	–	ct	837	835	801
8	–	mc	411	410	373

	Other software	
	Name	Time
9	PETfold	2519
10	RNAalifold	30

The labels 'pf', 'al', 'ct' and 'mc' indicate Pfold, RNAalipffold, CONTRAfold and McCaskill models, respectively. $p(\theta|A)$ and $p(\theta|x)$ are the probability distributions that correspond to the ones in Equation (2). The column 'time(all)' means the computational time for computing all the common secondary structures with the 17 γ-parameters of our data-set in order to obtain the SEN-PPV curve after computing the base-pairing probability matrices). The column 'time(MCC)' means the computational time for predicting the secondary structure with the pseudo-expected MCC. The column 'time(fixed)' means the computational time for computing a secondary structure with a fixed γ. The 5th, 6th, 7th and 8th rows are equivalent to Pfold-Centroid, RNAalipffold-Centroid, CentroidAlifold(old) with the CONTRAfold model and CentroidAlifold(old) with the McCaskill model, respectively.

proposed by Ding and Lawrence (45), and it has already been implemented in the software, CentroidAlifold. It should be noted that, if we use the sampling method, the credibility limit can be computed easily. Detailed study about the credibility limit for common secondary structures will be included in our future work.

CONCLUSION

In this study, we systematically discussed state-of-the-art algorithms of predicating secondary structure for aligned RNA sequences, and a classification of those algorithms were presented. Then, we introduced an improvement of CentroidAlifold, which was previously proposed by our group (17). Computational experiments have indicated

Table 4. SEN, PPV and MCC for CentroidAlifold, RNAalifold and PETfold with respect to Evaluation Process 2

Alignment	CentroidAlifold			PETfold			RNAalifold		
	SEN	PPV	MCC	SEN	PPV	MCC	SEN	PPV	MCC
Reference	0.79	**0.88**	**0.83**	**0.80**	0.80	0.80	0.73	0.82	0.78
ClustalW	0.43	**0.65**	**0.53**	**0.44**	0.59	0.51	0.36	0.67	0.49
ProbCons	0.54	**0.79**	**0.65**	**0.57**	0.66	0.61	0.45	0.75	0.58
MAFFT	0.59	**0.75**	**0.66**	**0.62**	0.66	0.64	0.54	0.72	0.62
MXSCARNA	0.64	**0.73**	**0.68**	**0.66**	0.66	0.66	0.52	0.74	0.62

In CentroidAlifold, we used a mixture of the probability distributions of the RNAalipffold model and the McCaskill model with the same weight, and selected the secondary structure using the pseudo-expected MCC. The bold values indicate the best values among three tools for each accuracy measure.

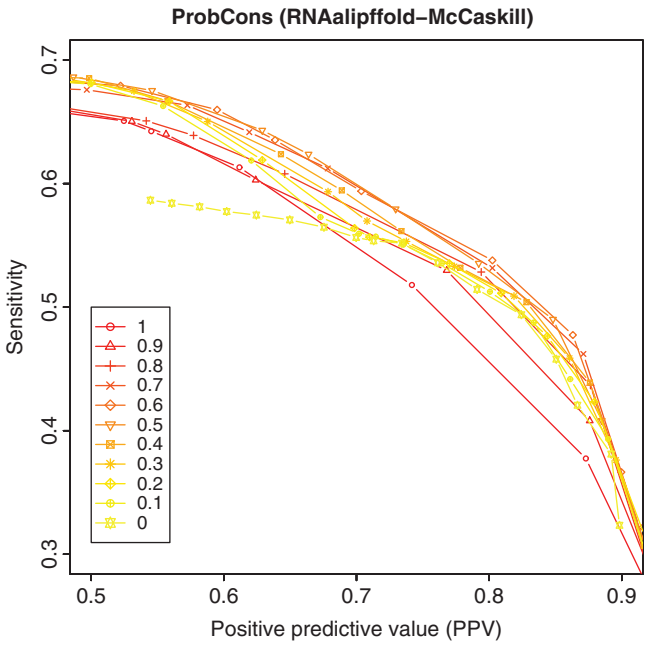

Figure 5. The performances of CentroidAlifold with various values of the weight parameter [i.e. $w = 0,0.1,0.2,\ldots,0.9,1$ in Equation (2)]. In this experiment, we used the mixture distribution of RNAalipffold model (12) and McCaskill model (18), and the alignments produced by ProbCons (34). The curves with $w = 1$ and $w = 0$ are equivalent to the 'previous' CentroidAlifold and RNAalipffold-Centroid, respectively. The results of the other combinations of probability distributions and aligners are shown in the Supplementary Data (Supplementary Figures S6–S10).

that the improved CentroidAlifold substantially outperformed the previous one and state-of-the-art algorithms, such as PETfold and RNAalifold. The software is freely available from web site: http://www.ncrna.org/software/centroidalifold, which will be useful for finding non-coding RNAs from genomic sequences or phylogenetic analyses of RNAs.

SUPPLEMENTARY DATA

Supplementary Data is available from NAR online.

ACKNOWLEDGEMENTS

The authors thank Drs/Profs Luis E. Carvalho, Charles E. Lawrence, Koji Tsuda, Hisanori Kiryu and Toutai Mituyama for useful discussions. We are grateful to Dr. Martin C. Frith for commenting on the manuscript. We are also grateful to the members of the Computational Biology Research Center (CBRC), the National Institute of Advanced Industrial Science and Technology (AIST).

FUNDING

'Functional RNA Project' of the New Energy Technology Development Organization (NEDO), Grant-in-Aid for Scientific Research on Innovative Areas (in parts). Funding for open access charge: Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST).

Conflict of interest statement. None declared.

REFERENCES

1. Bernhart,S.H. and Hofacker,I.L. (2009) From consensus structure prediction to RNA gene finding. *Brief. Funct. Genomic Proteomic*, **8**, 461–471.
2. Schroeder,S.J. (2009) Advances in RNA structure prediction from sequence: new tools for generating hypotheses about viral RNA structure-function relationships. *J. Virol.*, **83**, 6326–6334.
3. Hofacker,I., Fontana,W., Stadler,P., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
4. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
5. Clyde,K. and Harris,E. (2006) RNA secondary structure in the coding region of dengue virus type 2 directs translation start codon selection and is required for viral replication. *J. Virol.*, **80**, 2170–2182.
6. Jochl,C., Rederstorff,M., Hertel,J., Stadler,P.F., Hofacker,I.L., Schrettl,M., Haas,H. and Huttenhofer,A. (2008) Small ncRNA transcriptome analysis from Aspergillus fumigatus suggests a novel mechanism for regulation of protein synthesis. *Nucleic Acids Res.*, **36**, 2677–2689.
7. Okada,Y., Sato,K. and Sakakibara,Y. (2010) Improvement of structure conservation index with centroid estimators. In *Proceedings of the 15th Pacific Symposium on Biocomputing*. pp. 88–97.
8. Stocsits,R.R., Letsch,H., Hertel,J., Misof,B. and Stadler,P.F. (2009) Accurate and efficient reconstruction of deep phylogenies from structured RNAs. *Nucleic Acids Res.*, **37**, 6184–6193.
9. Thurner,C., Witwer,C., Hofacker,I.L. and Stadler,P.F. (2004) Conserved RNA secondary structures in Flaviviridae genomes. *J. Gen. Virol.*, **85**, 1113–1124.
10. Washietl,S., Hofacker,I.L., Lukasser,M., Huttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.
11. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
12. Bernhart,S., Hofacker,I., Will,S., Gruber,A. and Stadler,P. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
13. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
14. Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.
15. Seemann,S., Gorodkin,J. and Backofen,R. (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, **36**, 6355–6362.
16. Kiryu,H., Kin,T. and Asai,K. (2007) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics*, **23**, 434–441.
17. Hamada,M., Kiryu,H., Sato,K., Mituyama,T. and Asai,K. (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465–473.
18. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
19. Do,C., Woods,D. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
20. Sato,K., Hamada,M., Asai,K. and Mituyama,T. (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res.*, **37**, W277–W280.
21. Hamada,M., Sato,K., Kiryu,H., Mituyama,T. and Asai,K. (2009) Predictions of RNA secondary structure by combining homologous sequence information. *Bioinformatics*, **25**, i330–i338.
22. Lu,Z.J., Gloor,J.W. and Mathews,D.H. (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, **15**, 1805–1813.
23. Bradley,R.K., Pachter,L. and Holmes,I. (2008) Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics*, **24**, 2677–2683.
24. Bradley,R.K., Roberts,A., Smoot,M., Juvekar,S., Do,J., Dewey,C., Holmes,I. and Pachter,L. (2009) Fast statistical alignment. *PLoS Comput. Biol.*, **5**, e1000392.
25. Holmes,I. and Durbin,R. (1998) Dynamic programming alignment accuracy. *J. Comput. Biol.*, **5**, 493–504.
26. Sahraeian,S.M. and Yoon,B.J. (2010) PicXAA: greedy probabilistic construction of maximum expected accuracy alignment of multiple sequences. *Nucleic Acids Res.*, **38**, 4917–4928.
27. Frith,M.C., Hamada,M. and Horton,P. (2010) Parameters for accurate genome alignment. *BMC Bioinformatics*, **11**, 80.
28. Kall,L., Krogh,A. and Sonnhammer,E.L. (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, **21(Suppl. 1)**, i251–i257.
29. Michal,N., Tomas,V. and Brona,B. (2010) The highest expected reward decoding for hmms with application to recombination detection. *arXiv:1001.4499v1*, 2010 [Epub ahead of print, 25 Jan 2010].
30. Gross,S., Do,C., Sirota,M. and Batzoglou,S. (2007) CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol.*, **8**, R269, http://genomebiology.com/2007/8/12/R269.
31. Kato,Y., Sato,K., Hamada,M., Watanabe,Y., Asai,K. and Akutsu,T. (2009) RactIP: fast accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics*, 2009.
32. Hamada,M., Sato,K., Kiryu,H., Mituyama,T. and Asai,K. (2009) CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics*, **25**, 3236–3243.
33. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
34. Do,C., Mahabhashyam,M., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
35. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S. and Eddy,S.R. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
36. Andronescu,M., Bereg,V., Hoos,H. and Condon,A. (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.
37. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33(Database issue)**, 121–124.

38. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
39. Tabei,Y., Kiryu,H., Kin,T. and Asai,K. (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.
40. Carvalho,L. and Lawrence,C. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl Acad. Sci. USA*, **105**, 3209–3214.
41. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University press, Cambridge, UK.
42. Nussinov,R., Pieczenk,G., Griggs,J. and Kleitman,D. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math.*, **35**, 68–82.
43. Newberg,L.A. and Lawrence,C.E. (2009) Exact calculation of distributions on integers, with application to sequence alignment. *J. Comput. Biol.*, **16**, 1–18.
44. Webb-Robertson,B.J., McCue,L.A. and Lawrence,C.E. (2008) Measuring global credibility with application to local sequence alignment. *PLoS Comput. Biol.*, **4**, e1000077.
45. Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.