BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 4 2007, pages 434–441
doi:10.1093/bioinformatics/btl636

Structural bioinformatics

Robust prediction of consensus secondary structures using averaged base pairing probability matrices

Hisanori Kiryu[1,2,*], Taishin Kin[1] and Kiyoshi Asai[1,3]

[1]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan, [2]Graduate School of Information Sciences, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan and [3]Department of Computational Biology, Faculty of Frontier Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan

Received on September 9, 2006; revised on December 11, 2006; accepted on December 12, 2006
Advance Access publication December 20, 2006
Associate Editor: Charlie Hodgman

ABSTRACT

Motivation: Recent transcriptomic studies have revealed the existence of a considerable number of non-protein-coding RNA transcripts in higher eukaryotic cells. To investigate the functional roles of these transcripts, it is of great interest to find conserved secondary structures from multiple alignments on a genomic scale. Since multiple alignments are often created using alignment programs that neglect the special conservation patterns of RNA secondary structures for computational efficiency, alignment failures can cause potential risks of overlooking conserved stem structures.

Results: We investigated the dependence of the accuracy of secondary structure prediction on the quality of alignments. We compared three algorithms that maximize the expected accuracy of secondary structures as well as other frequently used algorithms. We found that one of our algorithms, called McCaskill-MEA, was more robust against alignment failures than others. The McCaskill-MEA method first computes the base pairing probability matrices for all the sequences in the alignment and then obtains the base pairing probability matrix of the alignment by averaging over these matrices. The consensus secondary structure is predicted from this matrix such that the expected accuracy of the prediction is maximized. We show that the McCaskill-MEA method performs better than other methods, particularly when the alignment quality is low and when the alignment consists of many sequences. Our model has a parameter that controls the sensitivity and specificity of predictions. We discussed the uses of that parameter for multi-step screening procedures to search for conserved secondary structures and for assigning confidence values to the predicted base pairs.

Availability: The C++ source code that implements the McCaskill-MEA algorithm and the test dataset used in this paper are available at http://www.ncrna.org/papers/McCaskillMEA/

Contact: kiryu-h@aist.go.jp

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

1 INTRODUCTION

Recently, a number of studies have shown that there is a substantial number of RNA transcripts that do not code protein sequences in higher eukaryotic cells (Okazaki *et al*., 2002; Dunham *et al*., 2004; Carninci *et al*., 2005), and the question whether such transcripts have any functional roles in cellular processes has attracted much interest. Since the existence of conserved secondary structures among phylogenetic relatives indicates the functional importance of such transcripts, several research groups have sought for conserved secondary structures on a genomic scale (Hofacker *et al*., 2004b; Washietl *et al*., 2005a,b; Pedersen *et al*., 2006).

In their studies, a large number of multiple alignments were created using computer programs, and consensus secondary structures were then predicted from these alignments. They used alignment programs that neglected the special conservation patterns of secondary structures, such as the base covariations in the stem regions since the alignment algorithms that took into account the base covariations required huge computational resources (Sankoff, 1985; Hofacker *et al*., 2004a; Mathews and Turner, 2002). Therefore, there were potential risks of overlooking conserved secondary structures due to misalignments. Such loss of sensitivity is particularly problematic in the early stage of large-scale screening that precedes the time-consuming but accurate computational and experimental validation stages.

In this paper, we investigate the dependence of the accuracy of secondary structure prediction on the quality of alignments and propose a method to predict conserved secondary structures from multiple alignments, which is robust against alignment failures. Our algorithm first computes the base pairing probability matrix for each sequence in the alignment and then obtains the base pairing probability matrix of the alignment by averaging over these matrices. The consensus secondary structure is predicted from this matrix using a Nussinov-style dynamic programming algorithm (Nussnov *et al*., 1978).

The use of the average pair probability matrix for obtaining the consensus structures is not a new idea and there have been several studies (Hofacker and Stadler, 1999; Hofacker *et al*., 1998; Luck *et al*., 1996, 1999; Knight *et al*., 2004) that have used the base pairing probability matrices of single sequences to predict the consensus secondary structures. In particular, the ConStruct program (Luck *et al*., 1996, 1999) predicts the same consensus structures as those predicted by our algorithm with a specific parameter value. However, we present a new interpretation and justification of this method in terms of the maximal expected accuracy (MEA) principle

[*]To whom correspondence should be addressed.

(Miyazawa, 1995), which has been successfully applied recently to the sequence alignment and the structure prediction to single sequences. This new interpretation makes it obvious that the method has an advantage in predicting the structures from seriously misaligned sequences. We show that our algorithm outperforms the leading programs for the consensus structure prediction (Hofacker *et al.*, 2002; Knudsen and Hein, 2003) in such a situation.

2 SYSTEMS AND METHODS

2.1 Definitions

We first present a few formal definitions of conserved secondary structure prediction that are useful for subsequent discussions. For a given alignment of length L, which is composed of N sequences X, let $\mathcal{C} = \{i \mid 1 \le i \le L\}$ be the set of positions of alignment columns and let $\mathcal{PC} = \{(i,j) \in \mathcal{C} \times \mathcal{C} \mid 1 \le i < j \le L\}$ be the set of pairs of alignment columns. A secondary structure of an alignment is defined by the mapping m from $\mathcal{PC}$ to the binary values $\{0,1\}$,

$$m : PC \to \{0,1\},$$

such that $m(i,j) = 1$ if the column pair (i,j) forms a base pair and 0 otherwise. We let $y = \{y_{ij} \in \{0,1\} \mid (i,j) \in \mathcal{PC}, y_{ij} = m(i,j)\}$ be the image of the mapping. y cannot assume all possible $2^{[L(L-1)/2]}$ values to form a consistent secondary structure. Since each column cannot be paired with two or more columns, they satisfy the following constraint:

$$\begin{aligned} &\exists(i,j), y_{ij} = 1 \\ &\Rightarrow \forall k \ne i,j, y_{ik} = y_{kj} = 0. \end{aligned} \tag{1}$$

Moreover, since we do not consider pseudo knot structures in this paper, we assume that y follows the nested structure constraint.

$$\begin{aligned} &\exists(i,j), y_{ij} = 1 \\ &\Rightarrow \forall(k,l), i < k < j < l \text{ or } k < i < l < j, y_{kl} = 0. \end{aligned} \tag{2}$$

For each sequence $x \in X$ in the alignment, we similarly consider the secondary structure $y^{(x)} = \{y_{ij}^{(x)} \mid (i,j) \in \mathcal{PC}\}$ of sequence x. $y_{ij}^{(x)}$ is always set to zero if the alignment column i or j is a gap position for the sequence x. We also use an alternative representation $\mathcal{S}$ of a consensus secondary structure that consists of a set of loop columns $\mathcal{L}$ and a set of pair columns $\mathcal{P}$.

$$\begin{aligned} \mathcal{S} &= \{\mathcal{L}, \mathcal{P}\} \\ \mathcal{L} &= \{i \in \mathcal{C} \mid \forall k \ne i, y_{ik} = y_{ki} = 0\} \\ \mathcal{P} &= \{(i,j) \in \mathcal{PC} \mid y_{ij} = 1\} \end{aligned}$$

2.2 Maximal expected accuracy algorithm

Recent studies have shown that the secondary structure predictions based on the principle of the MEA (Miyazawa, 1995) perform better than the predictions made by the conventional maximal likelihood algorithm (Pedersen *et al.*, 2006; Do *et al.*, 2006; Knudsen and Hein, 2003). This algorithm first computes the base pairing probability p_{ij} for each pair of alignment columns (i, j) and then considers the expected accuracy EA($\mathcal{S}$) for each secondary structure candidate $\mathcal{S}$. The predicted secondary structure $\mathcal{S}$ is obtained by maximizing the expected accuracy EA($\mathcal{S}$) with respect to $\mathcal{S}$. We first consider the secondary structure prediction for single sequences. For a given conditional probability distribution $P(y \mid x)$, the base pairing probability p_{ij} of columns (i,j) can be defined as follows.

$$\begin{aligned} p_{ij} &= E[\delta(y_{ij}, 1)] \\ &= \sum_{y} \delta(y_{ij}, 1) P(y \mid x) \\ &= \sum_{y \mid y_{ij}=1} P(y \mid x). \end{aligned}$$

Here, $\delta(z, z')$ is the Kronecker delta function and is defined by the following equation.

$$\delta(z, z') = \begin{cases} 1 & \text{if } z = z' \\ 0 & \text{otherwise} \end{cases}.$$

Further, $E[A]$ is the expected value of A with respect to $P(y \mid x)$. Let the loop probability q_i be the probability that the alignment column i does not form any pair with other columns.

$$\begin{aligned} q_i &= E\left[\prod_{j \ne i} \delta(y_{ij}, 0)\right] \\ &= E\left[\prod_{j \ne i} (1 - \delta(y_{ij}, 1))\right] \\ &= E\left[1 - \sum_{j \ne i} \delta(y_{ij}, 1)\right] \\ &= 1 - \sum_{j \ne i} p_{ij}. \end{aligned} \tag{3}$$

Here, we have used the convention $y_{ij} = y_{ji}$ and $p_{ij} = p_{ji}$ for $i > j$ to simplify the notation. q_i always assumes a non-negative value. In the third equation, we have used the constraint of Equation (1). For a secondary structure $\mathcal{S} = \{\mathcal{L}, \mathcal{P}\}$ and a given parameter $\alpha \ge 0$, the expected accuracy $\text{EA}_\alpha(\mathcal{S})$ of $\mathcal{S}$ with respect to the conditional distribution $P(y \mid x)$ is defined as follows.

$$\begin{aligned} \text{EA}_\alpha(\mathcal{S}) &= E\left[\alpha \sum_{i \in \mathcal{L}} \prod_{j \ne i} \delta(y_{ij}, 0) + 2 \sum_{(i,j) \in \mathcal{P}} \delta(y_{ij}, 1)\right] \\ &= \alpha \sum_{i \in \mathcal{L}} q_i + 2 \sum_{(i,j) \in \mathcal{P}} p_{ij}. \end{aligned}$$

When $\alpha = 1$, $\text{EA}_\alpha(\mathcal{S})$ can be interpreted as the expected value of the number of correctly annotated bases with respect to the conditional probability distribution $P(y \mid x)$.

The secondary structure that maximizes the expected accuracy can be computed by the traceback procedure of a Nussinov-like dynamic programming algorithm (Nussnov *et al.*, 1978).

$$M_{i,j} = \max \begin{cases} M_{i-1,j-1} + 2p_{ij} \\ M_{i-1,j} + \alpha q_i \\ M_{i,j-1} + \alpha q_j \\ M_{i,k} + M_{k+1,j} \text{ for } i < k < j. \end{cases} \tag{4}$$

The maximum of the expected accuracy MEA_α is given by the following equation.

$$\begin{aligned} \text{MEA}_\alpha &= \max_{\mathcal{S}} \text{EA}_\alpha(\mathcal{S}) \\ &= M_{1,L}. \end{aligned}$$

The corresponding secondary structure $\mathcal{S} = \mathcal{S}_{\text{MEA}}$ is the MEA solution. The parameter α controls the sensitivity and specificity of the structure prediction (Do *et al.*, 2006). A small α value encourages the base pair formation, which results in higher sensitivity, and a large α value encourages the increase of single-stranded regions and results in higher specificity.

$P(y \mid x)$ can be computed for various models, such as models based on the loop decomposition of secondary structure energy and models based on stochastic context-free grammars (SCFGs). In energy based models, $P(y \mid x)$ is given by the Boltzmann distribution of secondary structure configurations,

$$\begin{aligned} P(y \mid x) &= \frac{1}{Z(x)} \exp\left(-\frac{E(y,x)}{kT}\right) \\ Z(x) &= \sum_{y} \exp\left(-\frac{E(y,x)}{kT}\right), \end{aligned} \tag{5}$$

where $E(y, x)$ denotes the secondary structure energy that is computed using the energy parameters collected by the Turner group (Mathews *et al*., 1999). k is the Boltzmann constant, T is the temperature and $Z(x)$ is the partition function. In this case, the corresponding base pairing probability matrix $p = \{p_{ij} \mid (i,j) \in \mathcal{PC}\}$ is computed by McCaskill's algorithm (McCaskill, 1990).

The maximal likelihood prediction of the energy-based models is given by the secondary structure y that maximizes $P(y \mid x)$.

$$y = \mathrm{argmax}_{y'} P(y' \mid x)$$

The Mfold algorithm (Washietl and Hofacker, 2004; Zuder, 1989) is interpreted to be one such algorithm.

In the SCFG models (Dowell and Eddy, 2004), $P(y \mid x)$ is given by the sum of the conditional probabilities $P(\sigma \mid x)$ over the set of parses σ sharing the same secondary structure y.

$$\begin{aligned} P(y \mid x) &= \sum_{\sigma \in y} P(\sigma \mid x) \\ &= \frac{\sum_{\sigma \in y} P(\sigma, x)}{\sum_{\sigma} P(\sigma, x)}. \end{aligned}$$

$P(\sigma, x)$ 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. The sum of the numerator is over all the parse trees that share the same secondary structure y, and the sum of the denominator is over all the possible parse trees. The corresponding base pairing probability matrix is computed by the inside and outside algorithm (Durbin *et al*., 1998; Do *et al*., 2006).

The reason why the MEA algorithms show better prediction accuracy than their maximal likelihood counterparts can be explained as follows. In general, any computational model of secondary structure prediction based only on the sequence data has limitations in accuracy because an RNA molecule in reality forms a 3D structure in the cell, and interacts with itself among all the neighboring bases in the 3D structure. Moreover, it interacts with bounded proteins and other cellular environments that affect the formation of its secondary structure. Hence, the absolute optimality with respect to the scoring system of the model is of limited importance. When the model is not very accurate but reasonably good, taking the majority of near-optimal structures may be a more feasible way to predict the secondary structures. The MEA algorithm can be considered as one of such algorithms. For example, if an optimal structure does not form a pair at a column pair (i,j) but many suboptimal structures form a pair at (i,j), then the MEA solution tends to predict the pair (i,j) since the base pair probability p_{ij} assumes a large value at that position. Therefore, in contrast to the maximal likelihood method that considers only one optimal structure, the MEA algorithm takes into account various near-optimal structures and predicts the consensus structure supported by them. It presumably acts to reduce model-specific artifacts in the predictions.

A more pragmatic reason for the efficiency of the MEA algorithms is that the objective function of MEA is closer to the accuracy measures for the structure prediction. The accuracy of the structure prediction is usually evaluated by counting the (in-) correctly predicted base pairs, and not by mesuring the correctness of structural components, such as the loop lengths and the stacking energies. Such accuracy measures are advantageous to the MEA algorithm since the MEA solution tends to predict all the likely pairs throwing away the constraints of the original model that is used to compute the pair probability matrices.

2.3 Algorithms for consensus structure prediction

RNAAlipfold (Hofacker *et al*., 2002) is a multi-sequence extension of the McCaskill algorithm. For each consensus secondary structure candidate, it assigns a Boltzmann factor,

$$P(y \mid X) = \frac{1}{Z(X)} \exp\left(-\frac{E(y, X)}{kT} + \mathrm{Cov}(y, X) \right),$$

where $E(y, X)$ is the mean energy of the secondary structures of sequences X, all of which are assumed to form the same structure y, and $\mathrm{Cov}(y, X)$ is the base covariation bonus factor that gives a positive value for stem-conserving covariations. We consider an MEA algorithm that uses the base pairing probability matrices as calculated by RNAAlipfold and refer to it as RNAAlipfold-MEA. The maximal likelihood version of the RNAAlipfold algorithm corresponds to RNAAlifold (Hofacker *et al*., 2002) and is compared with other programs in the following section.

Pfold (Knudsen and Hein, 2003) is a multi-sequence extension of the SCFG model for a single sequence structure prediction. The differences from the single sequence case are that it simultaneously emits bases in each column and each pair of columns, and that the emission scores assume the likelihood values computed by the Markov model of sequence evolution. The Pfold algorithm is an MEA algorithm that maximizes the expected accuracy with $\alpha = 1$.

Both RNAAlipfold and Pfold assume the correctness of alignments, and the covariation scores contained in both the models rely on it. Their covariation scores are most efficient when they are applied to high-quality multiple alignments. However, in low-quality alignment data, there are many fake inconsistent mutations caused by alignment failures, which may cause the incorrect estimations of the covariation scores and result in the loss of sensitivity to conserved structures.

Here, we propose an alternative MEA algorithm that is not strongly dependent on the correctness of alignments. First, we define the conditional probability distribution function $P(Y \mid X)$ over all the secondary structures of all the sequences in the alignment as follows:

$$P(Y \mid X) = \prod_{x \in X} P(y^{(x)} | x),$$

where $Y = \{y^{(x)} \mid x \in X\}$ denote the set of secondary structures of sequences X, and $P(y^{(x)} \mid x)$ is given by the Boltzmann distribution of single sequence x [Equation (5)]. For each consensus structure candidate $\mathcal{S} = \{\mathcal{L}, \mathcal{P}\}$, we define the expected accuracy $\mathrm{EA}_\alpha(\mathcal{S})$ of the structure as the mean value of the expected accuracies of sequences.

$$\begin{aligned} \mathrm{EA}_\alpha(\mathcal{S}) &= E\left[\frac{1}{N} \sum_{x \in X} \left\{ \alpha \sum_{i \in \mathcal{L}} \prod_{j \neq i} \delta(y_{ij}^{(x)}, 0) + 2 \sum_{(i,j) \in \mathcal{P}} \delta(y_{ij}^{(x)}, 1) \right\} \right] \\ &= \alpha \sum_{i \in \mathcal{L}} q_i + 2 \sum_{(i,j) \in \mathcal{P}} p_{ij}, \end{aligned}$$

where p_{ij} is the mean value of the base pairing probabilities $p_{ij}^{(x)}$,

$$\begin{aligned} p_{ij} &= E\left[\frac{1}{N} \sum_{x' \in X} \delta(y_{ij}^{(x')}, 1) \right] \\ &= \sum_{y^{(x_1)}} \sum_{y^{(x_2)}} \cdots \sum_{y^{(x_N)}} \left(\frac{1}{N} \sum_{x' \in X} \delta(y_{ij}^{(x')}, 1) \right) P(Y \mid X) \\ &= \frac{1}{N} \sum_{x \in X} \sum_{y^{(x)}} \delta(y_{ij}^{(x)}, 1) P(y^{(x)} \mid x) \\ &= \frac{1}{N} \sum_{x \in X} p_{ij}^{(x)} \end{aligned} \tag{6}$$

and q_i is given by Equation (3).

The MEA structure is computed from q_i and p_{ij} in a manner identical to the case of the prediction from single sequences. We refer to the algorithm as McCaskill-MEA. For $\alpha = 0$, the McCaskill-MEA algorithm predicts the same structures as those of the ConStruct program (Luck *et al*., 1996, 1999).

The McCaskill-MEA algorithm does not assume that all the sequences take an equal single structure and instead predicts the structure that is supported by the majority of sequences. Since the model does not include any covariation score term, the accuracy of the prediction may be lower than that of other algorithms for high-quality alignments. However, McCaskill-MEA has the advantage that the algorithm is free from the negative effects of covariation scores in the presence of severe alignment errors.

To observe the effect of the suboptimal structures contained in the base pairing probability matrices $p_{ij}^{(x)}$, we consider another MEA algorithm.

Table 1. Test dataset used to compute ROC curves

Family name	Mean length	% identity	SPS (ProbCons)	SPS (ClustalW)
5S_rRNA	116	57	0.87	0.83
5_8S_rRNA	154	61	0.89	0.80
IRES_HCV	261	94	0.98	0.98
Lysine	181	49	0.76	0.61
RFN	140	66	0.91	0.84
Retroviral_psi	117	92	0.98	0.97
SECIS	64	41	0.68	0.35
SRP_bact	93	47	0.62	0.62
SRP_euk_arch	291	40	0.41	0.35
S_box	107	66	0.90	0.83
T-box	244	45	0.51	0.35
THI	105	55	0.84	0.58
U1	157	59	0.78	0.73
U2	182	62	0.76	0.67
UnaL2	54	73	0.97	0.85
sno_14q_I_II	75	64	0.93	0.83
tRNA	73	45	0.88	0.63
Average	142	59	0.80	0.70

For each family, we have collected five alignments of 10 sequences. In the first 3 columns, we have listed the family name, the mean length of sequences and the mean pairwise identity in percentage. In the last two columns, we listed the sum-of-pairs scores (SPS) of the alignment, which is generated by the ProbCons and ClustalW softwares. In the last row, the mean values of each column are shown. The dataset is used to compute ROC scores in Figures 2 and 3.

It first computes the predicted structures for all sequences using the Mfold program and defines the base pairing probability matrix of the sequence as

$$p_{ij}^{(x)} = \begin{cases} 1 & (i,j) \text{ is predicted to form a base pair by Mfold} \\ 0 & \text{otherwise} \end{cases}.$$

The base pairing probabilities of the alignment are computed as shown in Equation (6), and the predicted structure is computed by Equation (4). The algorithm is referred to as Mfold-MEA.

2.4 The dataset

We have collected a test dataset from version 7.0 of the Rfam database (Griffiths-Jones *et al*., 2003). We have used only the manually curated seed alignments with the consensus structures published in literatures. The alignments and structure annotations of the Rfam database is assumed to be correct and the accuracy is evaluated with respect to them. The test dataset consists of 85 subalignments of 10 sequences. The number of families is 17 and there are five subalignments for each family. The basic properties of the dataset are listed in Table 1, in which the family name, mean base length and mean sequence identity are listed.

The alignments have been realigned using the multiple alignment softwares ProbCons (Do *et al*., 2005) and ClustalW (Thompson *et al*., 1994). ProbCons is based on the MEA algorithm and its parameters are optimized for RNA sequences using the expectation maximization algorithm. ClustalW uses a simple heuristic scoring scheme that is not specifically tuned for RNA sequences. The accuracies of the alignments were evaluated by computing the sum-of-pairs score (SPS) (Carillo and Lipman, 1988). The obtained values are listed in Table 1. SPS is the fraction of base pairs in unequal sequences that are aligned in a manner identical to those in the reference alignment. SPS assumes values between zero and one.

Table 1 shows the diversity of the test dataset whose mean lengths vary from 54 bases to 291 bases and whose mean pairwise sequence identities vary from 40 to 94%. SPS also varies considerably among families; however, in general, SPS values of ProbCons alignments are higher than those of ClustalW alignments, indicating the superiority of ProbCons over ClustalW. The differences in the accuracy of the predictions between ProbCons alignments and ClustalW alignments can be considered as the differences in accuracy between high-quality alignments and low-quality alignments. Since the alignment quality is lower for multiple alignments of highly diverged sequences in general, the differences in accuracy between ProbCons and ClustalW alignments also indicate the differences in accuracy between the alignments of diverged sequences and those of evolutionarily close sequences.

The accuracy of structure predictions is evaluated by computing the sensitivity and specificity of the predictions, which are defined by

$$\text{specificity} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

$$\text{sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}},$$

where tp denotes the number of correctly predicted base pairs, fp denotes the number of incorrectly predicted base pairs and fn denotes the number of unpredicted true base pairs. The numbers are computed by assigning both the reference and predicted consensus structures to each sequence using the alignment and then counting the matches and mismatches of base pairs for all the sequences.

We also use the Matthews correlation coefficients (MCCs) (Matthews, 1975), which is defined by the formula.

$$\text{MCC} = \frac{\text{tp} \cdot \text{tn} - \text{fp} \cdot \text{fn}}{\sqrt{(\text{tp} + \text{fp})(\text{tp} + \text{fn})(\text{tn} + \text{fp})(\text{tn} + \text{fn})}}.$$

Here, tn denotes the number of correctly unpredicted base pairs.

3 IMPLEMENTATION

We used version 1.5 of the Vienna RNA package (Hofacker, 2003) for the computation of the base pairing probability matrices of McCaskill-MEA and RNAAlipfold-MEA and the structure predictions of the RNAAlifold algorithm. Version 5 of Mfold was used for the computation of the Mfold-MEA algorithm. A stand-alone program of Pfold was obtained (courtesy of Dr B. Knudsen). The algorithm of Equation (4) was implemented using the C++ language.

4 DISCUSSION

4.1 Comparison of algorithms

Figure 1 shows examples of the density plots of the base pairing probabilities, which are calculated from a multiple alignment of 10 tRNA sequences. The alignment is created using the ClustalW software. The lower left triangles in both the figures show the true distribution of base pair probabilities. They are computed by first assigning the annotated structure in the Rfam database to each sequence and then computing base pairing probability matrices in a manner similar to Mfold-MEA. Although, the true tRNA structure has only four stems (Fig. 1, bottom), about 10 stems are observed in the plot due to severe misalignments. The upper right triangles show the density plot of the pairing probabilities used in the RNAAlipfold-MEA (left) and McCaskill-MEA (right) algorithms. Only two out of four stems are observed in the matrix for RNAAlipfold-MEA, while all the four stems are observed in the matrix for McCaskill-MEA.

Figure 2 shows the receiver operator characteristic (ROC) curves of the structure predictions from alignments of 10 sequences. The *x* and *y* axis represent the specificity and the sensitivity of predictions,

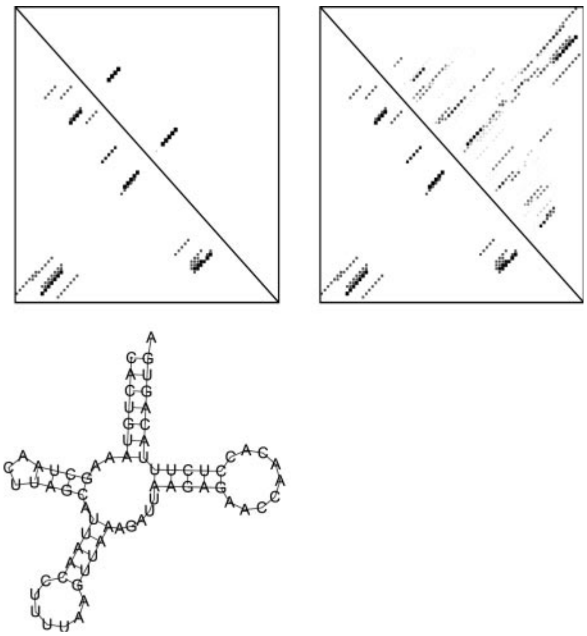

Fig. 1. Density plots of base pairing probability matrices. In both the matrices, the lower left triangle is the true distribution of base pair probabilities that is derived from the Rfam annotation of the tRNA secondary structure. The upper right triangles are the base pairing probabilities used in the RNAAlipfold-MEA and McCaskill-MEA algorithms, respectively. The true tRNA secondary structure is shown in the bottom figure, which is plotted using the RNAplot program of Vienna RNA package (Hofacker, 2003).

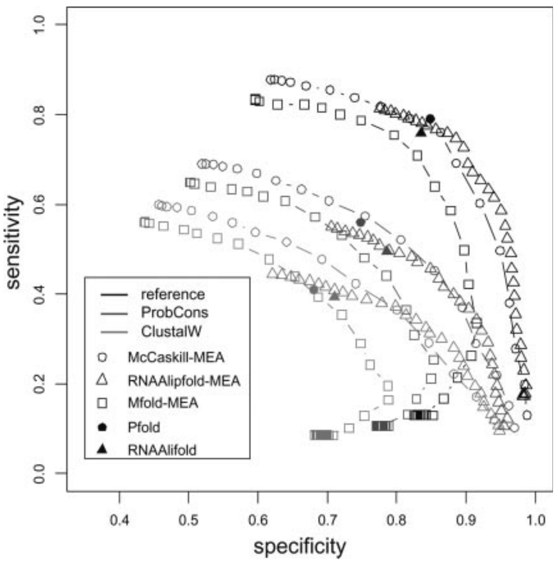

Fig. 2. ROC plot of the consensus structure predictions. The x and y axis represent the specificity and sensitivity of predictions, respectively. The colors indicate the types of alignments from which the consensus structures are predicted. The black, blue and red colors correspond to the reference alignments of the Rfam database, the ProbCons alignments and the ClustalW alignments, respectively. The character symbols indicate the types of structure prediction algorithms: McCaskill-MEA (open circle), RNAAlipfold-MEA (open triangle), Mfold-MEA (open square), Pfold (filled circle) and RNAAlifold (filled triangle). For McCaskill-MEA, RNAAlipfold-MEA and Mfold-MEA, multiple points are computed by varying the parameter α, and their trajectories are connected by lines. A colour version of this figure is available as supplementary data.

respectively. The ROC curves are computed by varying α in the three MEA algorithms. The sensitivity is large for small values of α, since the terms that score the base pairs in the expected accuracy is emphasized and the number of predicted base pairs increases. The ROC curve reaches a limit for $\alpha \to 0$. In this limit, the entire regions of the multiple alignments are filled with predicted stems. For large values of α, the number of predicted base pairs decreases. In the limit of large α, the number of predicted stems is so small that the corresponding plot fluctuates due to statistical fluctuations. Therefore, we only showed the data points for which the total number of predicted base pairs is greater than 10% of the total number of true base pairs. Since there is no parameter to control the specificity-sensitivity trade-off for Pfold (filled circle) and RNAAlifold (filled triangle), only one point for each alignment type is plotted.

Figure 2 shows that the sensitivity considerably depends on the alignment quality; the maximal sensitivity achieved for ClustalW alignments is less than ProbCons alignments by 10% and less than the reference alignments by about 30%. For all alignment types, the curves of McCaskill-MEA are above Mfold-MEA, which shows the efficiency achieved by including the effect of suboptimal structures in the consensus structure prediction. The higher sensitivity of RNAAlipfold-MEA as compared to that of RNAAlifold at the same specificity values also indicates the superiority of the MEA algorithm as compared to its maximal likelihood version, although the difference is less prominent. Both the specificity and sensitivity decrease for Mfold-MEA in the large α limit, which indicates that the loop probability values q_i incorrectly assume large values at the columns of the true base pairs due to the neglect of suboptimal structures. As expected, Pfold and RNAAlipfold-MEA show slightly better sensitivities for specificities >0.8 as compared to McCaskill-MEA due to the positive effect of the base covariation scores. However, McCaskill-MEA shows the best sensitivities for lower specificity regions in all the three alignment types.

Table 2 lists the ROC score, the maximal MCC and the maximal sensitivity for each alignment type and structure prediction algorithm. The ROC score is defined by the area under the ROC curve and is a standard indicator for prediction efficiency. Table 2 shows that the ROC score is the highest for the McCaskill-MEA algorithm. As for the maximal MCC, the Pfold program achieves the best MCC value for the high-quality reference alignments. However, McCaskill-MEA is better than the RNAAlifold program even for these alignments. For the ProbCons and ClustalW alignments, McCaskill-MEA algorithm outperforms the other programs. The difference between the maximal MCC value of McCaskill-MEA and that of other algorithms is larger for the lower quality ClustalW alignments. The table also shows that the maximal sensitivity is the highest for the McCaskill-MEA algorithm.

Note that in contrast to the specificity, the sensitivity values cannot be arbitrarily close to 1 unless the model's accuracy is fairly high; this is because it is not possible to predict all the base pair candidates (i.e. $y_{ij} = 1$ for all $1 \leq i \leq j \leq L$) to satisfy the consistency constraint [Equation (1)]. At this point, the

Table 2. The ROC score, maximal MCC and maximal sensitivity for each alignment type and structure prediction algorithm

Alignment	Algorithm	ROC score	Max MCC	Max sensitivity
Reference	McCaskill-MEA	**0.81**	0.81	**0.88**
	RNAAlipfold-MEA	0.77	0.81	0.81
	Mfold-MEA	0.73	0.77	0.83
	Pfold	—	**0.82**	0.79
	RNAAlifold	—	0.80	0.76
ProbCons	McCaskill-MEA	**0.60**	**0.66**	**0.69**
	RNAAlipfold-MEA	0.50	0.63	0.55
	Mfold-MEA	0.51	0.62	0.65
	Pfold	—	0.65	0.56
	RNAAlifold	—	0.62	0.50
ClustalW	McCaskill-MEA	**0.48**	**0.57**	**0.60**
	RNAAlipfold-MEA	0.39	0.54	0.44
	Mfold-MEA	0.40	0.54	0.56
	Pfold	—	0.53	0.41
	RNAAlifold	—	0.53	0.39

Since we can obtain only one point in the sensitivity-specificity plane for Pfold and RNAAlifold, we cannot show the ROC score for these softwares. Further, the maximal MCC and maximal sensitivity is the MCC and sensitivity at that point for these softwares. For other algorithms, the ROC score is defined as the area of the convex region that is spanned by the data points and the points $\{(0,0), (\mathrm{sp}_{\max}, 0), (0, \mathrm{sn}_{\max})\}$ in the specificity-sensitivity plane, where $\mathrm{sp}_{\max}$ and $\mathrm{sn}_{\max}$ denote the maximal specificity and sensitivity of the data points, respectively. The best scores for each alignment type are indicated in bold type face.

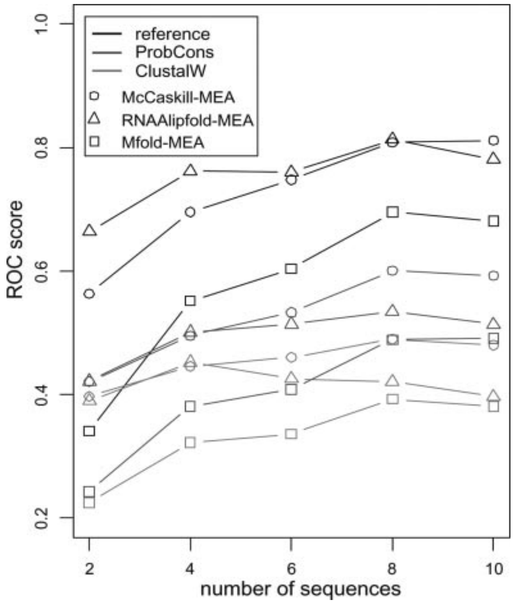

Fig. 3. Dependence of the ROC score on the number of sequences in the alignments. The colors and symbols have the same meanings as in Figure 2. A colour version of this figure is available as supplementary data.

secondary structure prediction problem is different from other binary classification problems where the classifier that predicts all the test samples as positive, (which corresponds to predicting $y_{ij} = 1$ for all $1 \leq i \leq j \leq L$), trivially achieves a sensitivity of one. Therefore, it may be said that the maximal reachable sensitivity is itself an indicator of the efficiency of the algorithms, and that McCaskill-MEA is comparatively much better than other algorithms with respect to it.

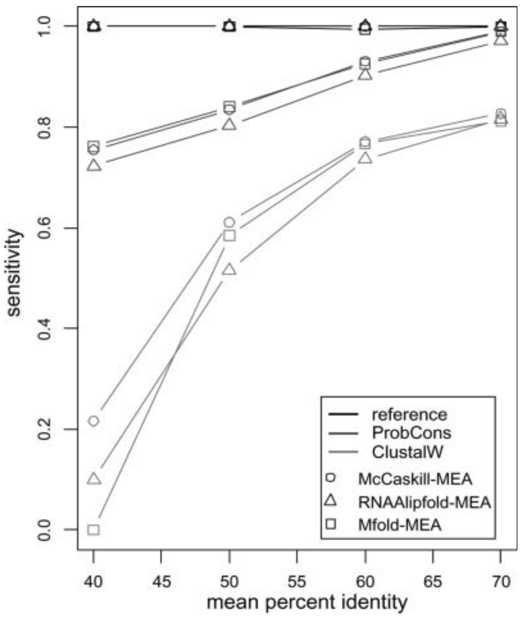

Fig. 4. An example of the sequence identity dependence of sensitivity at a specificity of 0.7. The colors and symbols have the same meanings as in Figures 2 and 3. The dataset is taken from the Hammerhead_3 family of the Rfam database. The dataset consists of 188 multiple alignments of four sequences. They are binned according to their mean pairwise sequence identities, and the result is averaged over each bin. A colour version of this figure is available as supplementary data.

Figure 3 shows the dependence of the ROC score on the number of sequences in the alignments. The alignments of 2, 4, 6 and 8 sequences are created by sampling sequences randomly from the alignments of 10 sequences. The colors and symbols are the same as in Figure 2. The ROC scores of McCaskill-MEA and Mfold-MEA increase with the number of sequences, while the increase of RNAAlipfold-MEA is somewhat slower than that of the other algorithms. For computationally aligned sequences, the McCaskill-MEA algorithm shows the best performance among the three algorithms. Even for the reference alignments, the McCaskill-MEA algorithm has a slightly better ROC score than the RNAAlipfold-MEA algorithm, which might imply the difficulty to score sequence covariations correctly for diverged sequences.

Figure 4 shows an example of the dependence of the sensitivity at fixed specificity (0.7) on the mean sequence identities. The current Rfam dataset has not permitted us to collect multiple sequence alignments over a wide-range of sequence identities for various families. The used dataset consists of 188 multiple alignments of four sequences collected from the Rfam Hammerhead_3 ribozyme family. The alignments are categorized into bins of width 10% in their mean sequence identities, and the mean sensitivities for each bin are plotted. The colors and symbols have the same meanings as in Figures 2 and 3. For this family, all three algorithms show similar behaviors for the reference alignments and the ProbCons alignments. For the ClustalW alignments, however, the McCaskill-MEA algorithm shows the best sensitivity in the region where the sequence identity is <60%.

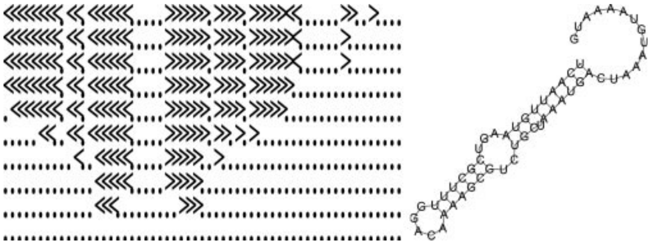

Fig. 5. An example of the consensus secondary structure prediction for varying the parameters α (left). A ProbCons alignment of the UnaL2 family in Table 1 is used. The predictions are made using the McCaskill-MEA algorithm. The corresponding α values are 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.63, 1.26, 2.51 and 5.01 from top to bottom. The true structure is also shown in the figure on the right.

4.2 Uses of parameter α

As we have shown, the prediction accuracy of conserved secondary structures significantly depends on the alignment quality, which indicates the necessity of refining the alignments after candidates of alignments with conserved structures are screened. The parameter α, which controls the sensitivity and specificity of the prediction accuracy, can be conveniently used for such multi-step screening procedures; this is done by taking small values of α to screen conserved structure candidates from coarse alignments with high-sensitivity and then taking α large to predict structures from refined multiple alignments with high-specificity.

Another use of the parameter α is to assign confidence values to predicted base pairs. Figure 5 shows an example of the predicted structures of the UnaL2 family for varying parameter α. They are predicted from a ProbCons alignment using the McCaskill-MEA algorithm. As seen from the figure, the number of base pairs monotonically decreases with α without creating alternative base pairs. This behavior holds in most cases. Hence, we define the confidence value of each predicted base pair as follows. For any base pair, we associate the α value that is maximal among the ones whose MEA solutions predict that pair, and define the confidence value of the pair as the specificity corresponding to that α [Fig. 6 (left)]. The confidence value of a predicted base pair represents the empirical probability that the pair is a true base pair. The definition of confidence value depends only on the test dataset and the corresponding ROC curve and is essentially independent of the details of models. It has the property that the number of base pairs with high-confidence values is large in the predicted structures from the accurate multiple alignments. The confidence values of the prediction to the UnaL2 family is plotted in Figure 6 (right). The confidence values will be useful to rank the secondary structure candidates in genomic scale studies of conserved secondary structures.

5 CONCLUSION

We have presented a method to predict the conserved secondary structures from multiple aligned sequences that are subject to alignment failures. Our method first calculates the base pairing probability matrix for each sequence, which are subsequently averaged to yield the base pairing probability matrix of the alignment. The consensus secondary structure is obtained by maximizing the expected accuracy of the structure with respect to the base pairing probabilities. For computationally aligned multiple sequences, our method shows a better performance as compared to other frequently used programs. We have shown that our method is particularly suitable for the alignments that suffer from significant alignment failures and that consist of a large number of sequences. We have shown that the parameter α in our model, which controls the sensitivity and specificity of the prediction, is useful for the genomic scale screening of conserved secondary structures and for assigning confidence values to the predicted base pairs.

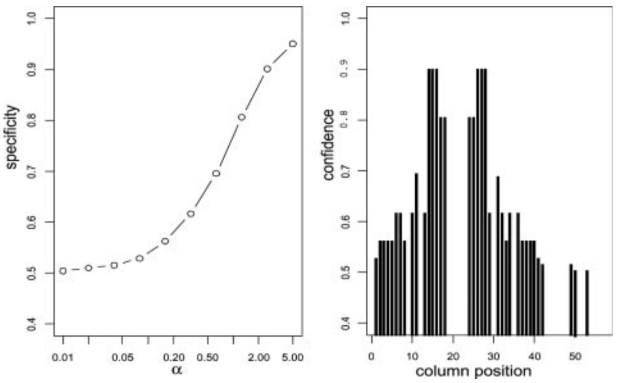

Fig. 6. Confidence scores of each predicted base pair. In the left-hand-side figure, the relation between the α values and specificity is plotted. The curve is derived from the ROC curve of the McCaskill-MEA predictions from the ProbCons alignments (Fig. 2). The x-axis denotes the α values on a log scale, and the y-axis denotes the specificity. In the right-hand-side figure, the confidence values of the predicted base pairs (Fig. 5) of the UnaL2 family are plotted. The x-axis indicates the column position of the alignment, and the y-axis indicates the computed confidence values.

In the present study, we have investigated only the global problem of structure prediction, i.e., the lengths of the alignments and those of the structural RNA genes are assumed to be of the same order. For the local problem of consensus structure prediction that searches long multiple alignments for small conserved structures, the calculation of the base pairing probability matrices over the entire alignment might have problems caused by the stochastic disturbance from the regions that are not related to the RNA genes and secondary structures. We leave the investigation of the local structure prediction problem and the scaling property of the base pairing probability matrices for future study.

We have considered only simple applications of the maximal expected accuracy principle in which the base pairing probabilities are derived either from an alignment or from all the sequences in the alignment. However, we can extend our method to compute the average of both the probabilities that might complement each other. We can also consider combining other probability matrices derived from other models, such as SCFG models. Such considerations lead to the problem of finding the best proportionality constants to sum the various probabilities. It may be interesting to study machine-learning approaches to combine various base pairing probabilities in an optimal manner.

ACKNOWLEDGEMENTS

The authors thank their colleagues in the project for discussions and comments. H.K. thanks Yasuo Tabei and Michiaki Hamada for helpful discussions. This work was partially supported by the ‘Functional

RNA Project' funded by the New Energy and Industrial Technology Development Organization (NEDO) of Japan.

Conflict of Interest: none declared.

REFERENCES

Carillo,H. and Lipman,D. (1988) The multiple sequence alignment problem in biology. *SIAM J. Appl. Math*., **48**, 1073–1082.

Carninci,P. *et al*. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.

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

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

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

Dunham,A. *et al*. (2004) The DNA sequence and analysis of human chromosome 13. *Nature*, **428**, 522–528.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G.J. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Griffiths-Jones,S. *et al*. (2003) Rfam: an RNA family database. *Nucleic Acids Res*., **31**, 439–441.

Hofacker,I. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*., **31**, 3429–3431.

Hofacker,I. and Stadler,P. (1999) Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput. Chem*., **23**, 401–414.

Hofacker,I. *et al*. (1998) Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res*., **26**, 3825–3836.

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

Hofacker,I. *et al*. (2004a) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.

Hofacker,I. *et al*. (2004b) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.

Knight,R. *et al*. (2004) BayesFold: rational 2 degrees folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences. *RNA*, **10**, 1323–1336.

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

Luck,R. *et al*. (1996) Thermodynamic prediction of conserved secondary structure: application to the RRE element of HIV, the tRNA-like element of CMV and the mRNA of prion protein. *J. Mol. Biol*., **258**, 813–826.

Luck,R. *et al*. (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res*., **27**, 4208–4217.

Mathews,D. and Turner,D. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol*., **317**, 191–203.

Mathews,D. *et al*. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol*., **288**, 911–940.

Matthews,B. (1975) Comparison of predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

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

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

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

Okazaki,Y. *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature*, **420**, 563–573.

Pedersen,J. *et al*. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol*., **2**, e33.

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

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

Washietl,S. and Hofacker,I. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol*., **342**, 19–30.

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

Washietl,S. *et al*. (2005b) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.

Zuker,M. (1989) Computer prediction of RNA structure. *Methods Enzymol*, **180**, 262–288.