BMC Bioinformatics

RESEARCH ARTICLE Open Access

Prediction of RNA secondary structure by maximizing pseudo-expected accuracy

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

Abstract

Background: Recent studies have revealed the importance of considering the entire distribution of possible secondary structures in RNA secondary structure predictions; therefore, a new type of estimator is proposed including the maximum expected accuracy (MEA) estimator. The MEA-based estimators have been designed to maximize the expected accuracy of the base-pairs and have achieved the highest level of accuracy. Those methods, however, do not give the single best prediction of the structure, but employ parameters to control the trade-off between the sensitivity and the positive predictive value (PPV). It is unclear what parameter value we should use, and even the well-trained default parameter value does not, in general, give the best result in popular accuracy measures to each RNA sequence.

Results: Instead of using the expected values of the popular accuracy measures for RNA secondary structure prediction, which is difficult to be calculated, the *pseudo*-expected accuracy, which can easily be computed from base-pairing probabilities, is introduced. It is shown that the pseudo-expected accuracy is a good approximation in terms of sensitivity, PPV, MCC, or F-score. The pseudo-expected accuracy can be approximately maximized for each RNA sequence by stochastic sampling. It is also shown that well-balanced secondary structures between sensitivity and PPV can be predicted with a small computational overhead by combining the pseudo-expected accuracy of MCC or F-score with the γ-centroid estimator.

Conclusions: This study gives not only a method for predicting the secondary structure that balances between sensitivity and PPV, but also a general method for approximately maximizing the (pseudo-)expected accuracy with respect to various evaluation measures including MCC and F-score.

Background

To predict the secondary structure of an RNA sequence is a classic problem of sequence analysis in bioinformatics. The importance of accurate predictions of secondary structures has increased due to the recent finding of functional non-coding RNAs whose functions are closely related to their secondary structures [1-3]. Secondary structure prediction also plays an important role in research on viral RNAs [4].

There are many tools and algorithms for secondary structure prediction [5-11]. The most popular approach is to predict the minimum free energy (MFE) structure by using the Zuker algorithm [12]. Well-known software (Mfold [13], RNAfold [14] and RNAstructure [15]) employs this approach. From a probabilistic viewpoint, the MFE structure is equivalent to the secondary structure of the maximum likelihood (ML) estimation for the probability distribution of secondary structures given by the McCaskill model [16]. It is, however, known that the MFE/ML structure has drawbacks: there are a huge number of suboptimal structures whose free energies are similar to the minimum free energy and the probability of the MFE structure is extremely small [17]. Moreover, the ML-estimator is not optimized for accuracy measures in the target problem [10].

Therefore, another approach that considers the entire distribution of possible secondary structures of a given sequence has been introduced. Ding *et al.* [18] proposed the centroid estimator, which minimizes the expected Hamming loss. On the other hand, Do *et al.* [7] proposed the maximum expected accuracy (MEA) estimator, which gives a prediction based on maximizing the

* Correspondence: hamada-michiaki@aist.go.jp
[1]Graduate School of Frontier Sciences, The University of Tokyo,5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan
Full list of author information is available at the end of the article

expected value of an accuracy function under a probabilistic distribution of secondary structures. The MEA-based estimators have been applied to many problems in bioinformatics, including sequence analyses for RNA sequences [6,7,10,19-22], alignment of bio- logical sequences [23-25] and other estimation problems [26-28].

For RNA secondary structure predictions, two MEA-based estimators have been proposed: (i) the estimator proposed by [7] and (ii) the γ-centroid estimator proposed by [10]. Both estimators do *not* employ the accuracy measures that are used in actual evaluation of RNA secondary structure, namely, sensitivity (SEN), positive predictive value (PPV), Matthews correlation coefficient (MCC) and F-score, with respect to predicted base-pairs. From a view- point of MEA, it is useful to consider the estimators that maximize expectation of those accuracy measures. Because the computation of those estimators generally demands huge computational time, the previous studies could not use those accuracy measures directly.

Moreover, the previous MEA-based estimators contain a parameter that controls the balance between SEN and PPV of base-pairs in a predicted secondary structure. It is, however, unclear how to select the parameter in order to obtain one reasonable secondary structure (e.g., a well-balanced secondary structure between SEN and PPV), although there are situations that only one predicted secondary structure is required. There is also a possibility that the optimal parameter might depend on the length of sequence and/or the type of RNA family, although the γ-centroid estimator (and the estimator proposed by [7]) employs a default parameter, determined by a benchmark dataset, which is identical for all sequences.

In this study, to address the drawbacks of the current MEA-based methods described above, We introduce the *pseudo*-expected accuracy of a secondary structure with respect to a given accuracy measure, which is a function of the number of true positive base-pairs (TP), true-negative base- pairs (TN), false-positive base-pairs (FP) and false- negative base-pairs (FN). The pseudo-expected accuracy is then defined by using expected TP, TN, FP and FN. As the accuracy measures, we utilize SEN, PPV, MCC and F-score with respect to base-pairs, which are commonly used in the evaluations of RNA secondary structure predictions, because the base- pairs are essential for forming secondary/tertiary structures, which are known to be biologically important.

The pseudo-expected accuracy is easily calculated using the base-pairing probability matrix, and can be computed much more efficiently than the expected accuracy. Although the pseudo-expected accuracy is not equal to the expected accuracy of a predicted secondary structure, we found that the pseudo-expected accuracy gives a good approximation of the expected accuracy in our situation. Accordingly, we also introduce the approximated estimators that maximize the expected accuracy of a given accuracy measure. Moreover, by combining the pseudo-expected MCC/F-score with the γ- centroid estimator, it is possible to predict the balanced secondary structure between SEN and PPV (which seems to be a reasonable secondary structure in many situations when only one predicted secondary structure is required), although there is a small computational overhead.

The techniques described in this paper will be extended to design the maximum expected accuracy estimator for various evaluation measures (cf. [29]).

Methods

In the following, we represent a secondary structure of an RNA sequence x as a triangular binary matrix:$\theta = \{\theta_{ij}\}_{i<j}$ where $\theta_{ij} = 1$ means that i-th base x_i and j-th base x_j form a base-pair, and $\theta_{ij} = 0$ means that i-th base x_i and j-th base x_j do not form a base-pair. In this study, pseudo-knotted base-pairs are not allowed in a secondary structure. For an RNA sequence $x, S(x)(\subset \{\theta_{ij} \in \{0,1\} \mid 1 \le i < j \le |x|\})$ denotes the space of all the possible secondary structures of x, where $|x|$ is the length of x. A probability distribution on $S(x)$ (denoted by $p(\cdot|x)$) is given by the McCaskill [16], CONTRAfold [7] or Simfold [11] models. The base-pairing probability matrix of x, $\{p_{i,j}\}_{i<j}$, has entries $p_{ij} = \sum_{\theta \in S(x)} I(\theta_{ij} = 1) p(\theta \mid x)$, called base-pairing probabilities, where $I(\cdot)$ is the indicator function that returns 1 if the condition is true and 0 otherwise. The base-pairing probability matrix of a given RNA sequence x can be computed using the McCaskill (Inside-Outside) algorithm, whose complexities are $O(|x|^3)$ and $O(|x|^2)$ for time and space, respectively (e.g., see [16,30]).

Expected accuracy and pseudo-expected accuracy of RNA secondary structure

Accuracy measures for RNA secondary structure prediction

For two secondary structures $\theta = \{\theta_{ij}\}_{i<j} \in S(x)$ and $\sigma = \{\sigma_{ij}\}_{i<j} \in S(x)$ of an RNA sequence x, we define

$$\mathrm{TP} = \mathrm{TP}(\theta, \sigma) = \sum_{i<j} I(\sigma_{ij} = 1) I(\theta_{ij} = 1), \tag{1}$$

$$\mathrm{TN} = \mathrm{TN}(\theta, \sigma) = \sum_{i<j} I(\sigma_{ij} = 0) I(\theta_{ij} = 0), \tag{2}$$

$$\mathrm{FP} = \mathrm{FP}(\theta,\sigma) = \sum_{i<j} I(\sigma_{ij} = 1) I(\theta_{ij} = 0), \tag{3}$$

$$\mathrm{FN} = \mathrm{FN}(\theta,\sigma) = \sum_{i<j} I(\sigma_{ij} = 0) I(\theta_{ij} = 1), \tag{4}$$

$$\mathrm{SEN} = \mathrm{SEN}(\theta,\sigma) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{5}$$

$$\mathrm{PPV} = \mathrm{PPV}(\theta,\sigma) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{6}$$

$$\mathrm{MCC} = \mathrm{MCC}(\theta,\sigma) = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{7}$$

$$\text{F-score} = \text{F-score}(\theta,\sigma) = \frac{2\cdot\mathrm{TP}}{2\cdot\mathrm{P} + \mathrm{FP} + \mathrm{FN}}. \tag{8}$$

When θ is a *reference* (correct) secondary structure and is a *predicted* secondary structure of x, Eqs. (1), (2), (3), (4), (5), (6), (7), and (8) are the number of true positive base-pairs, the number of true negative base-pairs, the number of false positive base-pairs, the number of false negative base-pairs, the SEN, the PPV, the MCC and the F-score, respectively. Because the base-pairs in a secondary structure are biologically important, accuracy measures based on base-pairs are useful and SEN, PPV, MCC and F-score are widely-used accuracy measures for secondary structure predictions. Note that MCC and F-score are balanced measures between SEN and PPV. (F-score is equal to a harmonic mean of SEN and PPV.) In the following, *Acc* means one of the SEN, PPV, MCC and F-score.

Expected accuracy of secondary structure

Given a probability distribution $p(\theta|x)$ on $S(x)$, we calculate the expected values of Eq. (1) to Eq. (4).

$$\widehat{\mathrm{TP}}(\sigma) = E_{\theta|x}[\mathrm{TP}(\theta,\sigma)] = \sum_{i<j} p_{ij} I(\sigma_{ij} = 1), \tag{9}$$

$$\widehat{\mathrm{TN}}(\sigma) = E_{\theta|x}[\mathrm{TN}(\theta,\sigma)] = \frac{|x|(|x|-1)}{2} - \sum_{i<j} I(\sigma_{ij} = 1) - \sum_{i<j} p_{ij} + \sum_{i<j} p_{ij} I(\sigma_{ij} = 1), \tag{10}$$

$$\widehat{\mathrm{FP}}(\sigma) = E_{\theta|x}[\mathrm{FP}(\theta,\sigma)] = \sum_{i<j} (1 - p_{ij}) I(\sigma_{ij} = 1), \tag{11}$$

$$\widehat{\mathrm{FN}}(\sigma) = E_{\theta|x}[\mathrm{FN}(\theta,\sigma)] = \sum_{i<j} p_{ij} (1 - I(\sigma_{ij} = 1)), \tag{12}$$

Where $\{p_{ij}\}$ indicates the base-pairing probability matrix. Moreover, we calculate the expected accuracy of an accuracy measure *Acc* (*Acc* is equal to SEN, PPV, MCC or F-score) of σ as follows:

$$\widehat{Acc}(\sigma) = E_{\theta|x}[Acc(\theta,\sigma)] = \sum_{\theta\in S(x)} Acc(\theta,\sigma) p(\theta \mid x). \tag{13}$$

In order to compute the expected *Acc* for a given secondary structure σ (i.e., $\widehat{Acc}(\sigma)$), it is necessary to sum over all the secondary structures of the RNA sequence x because no efficient algorithm (such as a dynamic programming algorithm) has been reported. The number of candidate secondary structures increases exponentially with the length of the RNA sequence (more precisely, there are roughly 1.8^L possible structures for a sequence of length L), so to compute the expected *Acc* is an intractable problem. Therefore, we approximate it using *stochastic sampling*: For N secondary structures $\{\theta^{(n)}\}_{n=1}^{N}$ given by stochastic sampling [30,31] of secondary structures, we define

$$\widehat{Acc}^{N}(\sigma) = \frac{1}{N} \sum_{1\le n\le N} Acc(\theta^{(n)},\sigma) \tag{14}$$

for $\sigma \in S(x)$. $\widehat{Acc}^{N}(\sigma)$ converges to $\widehat{Acc}(\sigma)$ when N is sufficiently large by the properties of stochastic sampling. It should be noted that the sample size N grows exponentially with the sequence length to $\widehat{Acc}^{N}(\sigma)$ be a reliable approximation to the expected *Acc* of σ.

Pseudo-expected accuracy of secondary structure

In our situation, *Acc* is generally written as a function of TP, TN, FP and FN:

$Acc = f(\mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \mathrm{FN})$

Then, for a secondary structure σ, the *pseudo*-expected *Acc* of is defined by

$$\widehat{Acc}^{0}(\sigma) = f(\widehat{\mathrm{TP}}, \widehat{\mathrm{TN}}, \widehat{\mathrm{FP}}, \widehat{\mathrm{FN}}). \tag{15}$$

For example, the pseudo-expected MCC is given by

$$\widehat{MCC}^{0}(\sigma) = \frac{\widehat{\mathrm{TP}}.\widehat{\mathrm{TN}} - \widehat{\mathrm{FP}}.\widehat{\mathrm{FN}}}{\sqrt{\left(\widehat{\mathrm{TP}}+\widehat{\mathrm{FP}}\right)\left(\widehat{\mathrm{TP}}+\widehat{\mathrm{FN}}\right)\left(\widehat{\mathrm{TN}}+\widehat{\mathrm{FP}}\right)\left(\widehat{\mathrm{TN}}+\widehat{\mathrm{FN}}\right)}}. \quad (16)$$

If we have the base-pairing probability matrix of x, the pseudo-expected *Acc* of σ can be easily computed by using Eqs. (9), (10), (11) and (12) without employing sampling/enumerating algorithms. Although the pseudo-expected *Acc* is *not* equal to the expected *Acc*, we shall see later that the pseudo-expected *Acc* is a good approximation of the expected *Acc* when *Acc* is equal to SEN, PPV, MCC or F-score.

Prediction of secondary structure by maximizing pseudo-expected accuracy

The γ-centroid estimator [10] for RNA secondary structure prediction is defined by

$$\hat{\sigma} = \arg\max_{\sigma \in S(x)} \left[\gamma \cdot \widehat{\mathrm{TP}}(\sigma) + \widehat{\mathrm{TN}}(\sigma) \right] \quad (17)$$

where γ >0 controls SEN and PPV of a predicted secondary structure. This estimator is suitable when we would like to predict more TP and TN and fewer FP and FN because Eq. (17) is equivalent to

$$\hat{\sigma} = \arg\max_{\sigma \in S(x)} \left[\alpha_1 \widehat{\mathrm{TP}}(\sigma) + \alpha_2 \widehat{\mathrm{TN}}(\sigma) - \alpha_3 \widehat{\mathrm{FP}}(\sigma) - \alpha_4 \widehat{\mathrm{FN}}(\sigma) \right] \quad (18)$$

with $\gamma = (\alpha_1 + \alpha_4)/(\alpha_2 + \alpha_3)$ and $\alpha_k \geq 0$. Hamada *et al*. [10] show that the secondary structure of the γ-centroid estimator can be calculated by Nussinovstyle dynamic programming.

Eq. (18) indicates that the γ-centroid estimator is not optimized for the actual evaluation measures (cf. SEN, PPV, MCC and F-score). It is useful to introduce the estimator that maximizes expected SEN, PPV, MCC or F-score directly:

$$\hat{\sigma} = \arg\max_{\sigma \in S(x)} \widehat{Acc}(\sigma). \quad (19)$$

It is, however, difficult to compute the expected *Acc* efficiently for given σ and $p(\theta|x)$. Because *Acc* contains products or divisions of TP, TN, FP and FN, no efficient method to compute the estimator Eq. (19) has been found, in contrast to the γ-centroid estimator of Eq. (17). Instead, we consider estimators that maximize *pseudo*-expected *Acc* as follows.

$$\hat{\sigma} = \arg\max_{\sigma \in S(x)} \widehat{Acc}^{0}(\sigma). \quad (20)$$

Prediction of secondary structure by maximizing pseudo-expected SEN/PPV

The pseudo-expected SEN and PPV of a secondary structure σ can be computed by

$$\widehat{\mathrm{SEN}}^{0}(\sigma) = \frac{\sum_{i<j} p_{ij} I(\sigma_{ij} = 1)}{\sum_{i<j} p_{ij}}, \quad (21)$$

$$\widehat{\mathrm{PPV}}^{0}(\sigma) = \frac{\sum_{i<j} p_{ij} I(\sigma_{ij} = 1)}{\sum_{i<j} I(\sigma_{ij} = 1)}. \quad (22)$$

Therefore, the secondary structure given by maximizing pseudo-expected SEN (Eq. (20) with SEN)) is equivalent to the secondary structure that maximizes the sum of base-paring probabilities of the predicted base-pairs. The secondary structure is, therefore, equivalent to the one given by the γ-centroid estimator with a sufficiently large γ [10]. On the other hand, the secondary structure given by maximizing pseudo-expected PPV (Eq. (20) with PPV)) is equivalent to the secondary structure that consists of (only) one base-pair that has the highest base-pairing probability. (The structure does not seem to be useful.) It should be noted that both structures can be easily computed by using the base-paring probability matrix of the target RNA sequence.

Prediction of secondary structure by maximizing pseudo-expected MCC/F-score with stochastic sampling (Method M1)

Because of the computational difficulty of computing "argmax" in Eq. (20) with MCC and F-score (see "Discussion" section for more details), we need to evaluate all the secondary structures in $S(x)$. The number of secondary structures of a given RNA sequence, however, is so large that it is not practical to enumerate all of them. Therefore, we again employ the stochastic sampling of secondary structures and approximate the estimator of Eq. (20) by

$$\hat{\sigma} = \arg\max_{\sigma \in S} \widehat{Acc}^{0}(\sigma). \quad (23)$$

where S is a set of secondary structures given by stochastic sampling. Note that the computational cost of this estimator is much smaller than that of predictions based on maximizing the expected MCC/F-score. If the pseudo-expected MCC/F-score gives a good approximation of the expected MCC/F-score and we use a sufficiently large sample size, then the estimator in Eq. (23) should be a reliable approximation to the estimator in Eq. (19) that maximizes the expected MCC/F-score.

Prediction of secondary structure with γ-centroid estimator and pseudo-expected MCC/F-score (Method M2)

In the γ-centroid estimator [10] of Eq. (17) implemented in the software CentroidFold [32], there is a parameter γ that adjusts the balance between SEN and PPV. It is, however, unclear how to select the γ parameter that achieves a reasonable structure although there are several situations that only one predicted secondary structure is required. As described in the previous section, we can predict the secondary structures that maximize (pseudo-)expected SEN or PPV, but the well-balanced secondary structure between SEN and PPV will be more useful in many cases than those structures.

Eq. (18), which is equivalent in form to the γ-centroid estimator, indicates that the γ-centroid estimator can *arbitrarily* control the number/ratio of the true predictions and the false predictions by using the parameter. By combining the pseudo-expected MCC/F-score with the γ-centroid estimator, it is possible to predict the balanced secondary structure between SEN and PPV, as follows. First, we compute the base-pairing probability matrix of the given RNA sequence, and then predict a set of secondary structures S^g of x by using the γ-centroid estimators with 17 γ parameters: $\gamma \in \{2^k: -5 \le k \le 10, k \in \mathbb{Z}\} \cup \{6\}$ that were used in our previous paper to obtain the SEN-PPV curve [7,10]. Here, the secondary structure of the γ-centroid estimator with $\gamma \in \{2^k: 0 < k \le 10, k \in \mathbb{Z}\} \cup \{6\}$ is computed by using Nussinov-style dynamic programming, but the secondary structure of the γ-centroid estimator with $\gamma \in \{2^k: -5 \le k \le 0, k \in \mathbb{Z}\}$ can be predicted *without* dynamic programming by selecting all the base-pairs whose probability is larger than 1/(γ+ 1) [10]. Second, we select the secondary structure in S^g that has the best pseudo-expected MCC/F-score:

$$\hat{\sigma} = \arg\max_{\sigma \in S^g} \widehat{Acc}^{0}(\sigma) \tag{24}$$

where *Acc* is equal to MCC or F-score. The pseudo-expected MCC/F-score of each secondary structure $\sigma \in S^g$ is easily computed because the base-pairing probability matrix has already been computed.

In this case, using the γ-centroid estimator is more suitable than using the MEA-based estimator proposed by [7], which also has a parameter that controls the balance between SEN and PPV, because the latter has a *bias* to MCC and F-score (see [10] for details).

Results

We conducted all experiments using a Linux OS machine, which has a 2 GHz AMD Opteron Model 246 processor and 4 GB of memory.

Experimental settings

For the dataset, we utilized the S-151Rfam dataset [7] that contains 151 RNA sequences with reference structures, each of which was taken from a different family in the Rfam database [1] This dataset has been widely used in previous studies of RNA secondary structure prediction, for example, [7,10,11]. For the probability distribution $p(\theta|x)$ of the secondary structures of RNA sequence x, we used the CONTRAfold model (version 2.02) [7] and the McCaskill model [16] (in the Vienna RNA pack-age version 1.8.3 [14]). For evaluation measures, we employed SEN, PPV, MCC and F-score with respect to the base-pairs, which are defined by Eqs. (5), (6), (7) and (8), respectively, where σ is a predicted structure and θ is a reference structure.

Comparison between pseudo-expected accuracy and expected accuracy

In this experiment, we compared the pseudo-expected *Acc* (Eq. (15)) with the expected *Acc* (Eq. (13)) where *Acc* is SEN, PPV, MCC or F-score. First, we obtained a set of secondary structures from the S-151Rfam dataset in the following way. For each RNA sequence in the S-151Rfam dataset, we predicted the secondary structures using the γ-centroid estimator [10] (implemented in CentroidFold) with 17 γ parameters, $\gamma \in \{2^k: -5 \le k \le 10\} \cup \{6\}$ and two models (the McCaskill [16] and CONTRAfold [7] models). Then, duplicate secondary structures were removed from the set. The set of the secondary structures contains various secondary structures, because the γ-centroid estimator with small γ predicts a small number of base-pairs and the one with large γ predicts a large number of base-pairs [10]. As described in the previous section, it is not feasible to compute the expected *Acc* (Eq. (13)) of a given secondary structure, because the number of possible secondary structures is immense. Therefore, we plotted $\widehat{Acc}^{0}(\sigma)$ (i.e., pseudo-expected *Acc* of; Eq. (15)) and $\widehat{Acc}^{1M}(\sigma)$ (i.e., expected *Acc* of σ approximated by 1 M (1,000,000) samples; Eq. (14)) for each secondary structure σ in the set of secondary structures.

The result is shown in Figure 1, which indicates the pseudo-expected SEN, PPV, MCC and F-score of the predicted secondary structure is a reliable approximation to the expected SEN, PPV, MCC and F-score, respectively. The averaged squared errors of the pseudo-expected SEN, PPV, MCC and F-score with respect to the CONTRAfold model and the McCaskill model are shown in Additional file 1, Table S1.

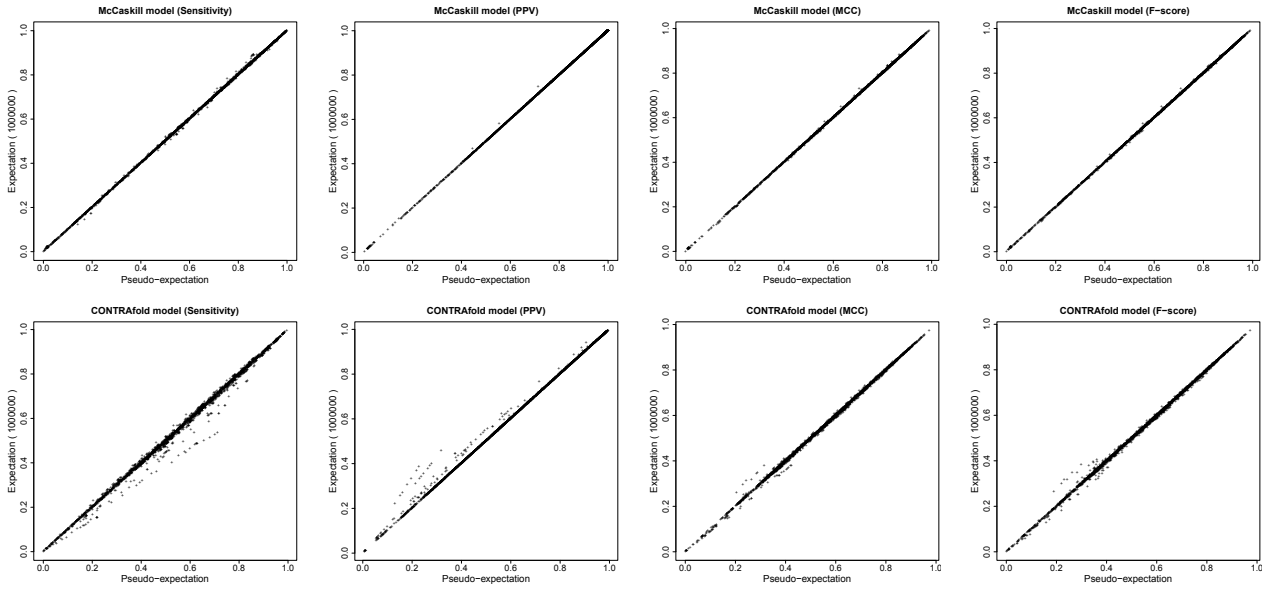

Figure 1 Comparison between pseudo-expected accuracy and expected accuracy. Comparison between the *pseudo*-expected SEN, PPV, MCC and F-score (the horizontal axes) and the expected SEN, PPV, MCC and F-score that are computed by stochastic sampling with a sample size of n = 1 M (the vertical axes). We used the McCaskill model (top row) and the CONTRAfold model (bottom row). The 1st, 2nd, 3rd and 4th columns indicate SEN, PPV, F-score and MCC, respectively. See Additional file 1, Figure S1 and Figure S2 for other sample sizes.

Results of secondary structure prediction by maximizing pseudo-expected accuracy

We conducted the experiments on RNA secondary structure prediction by maximizing the pseudo-expected MCC/F-score of the predicted secondary structure with stochastic sampling (the estimator in Eq. (23)). Note that the results in the previous section suggest that the estimator of Eq. (23) with a sufficiently large sample size is a good approximation to the estimator of Eq. (19) that maximizes the expected MCC/F-score.

The results are shown in Figure 2 (MCC) and Additional file 1, Figure S1 (F-score). As the sample size increased, the performance of the prediction of the estimator in Eq. (23) converged to the points on the SEN-PPV curves of the γ-centroid estimator [10], and favorable MCC/F-scores were achieved (Table 1). On the other hand, we need to sample a large number of secondary structures (more than 1 million) in order to obtain the secondary structure that has a good MCC/F-score. The computational time of the estimator of Eq. (23) increased linearly with the sample size (Table 2). The result also suggests that it is difficult to improve the performance of the γ-centroid estimator even if we employ the estimator of Eq. (19), that is, maximizing expected MCC/F-score.

It should be noted that the performance of the estimator that maximizes the pseudo-expected SEN (PPV) corresponds to the leftmost (rightmost) point in the SEN-PPV curve of the γ-centroid estimators.

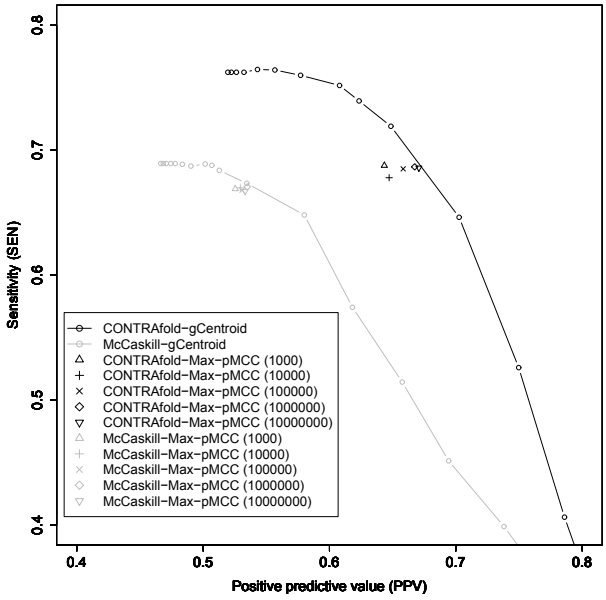

Figure 2 Performance of RNA secondary structure prediction by maximizing the pseudo-expected MCC with the stochastic sampling (Method M1). Performance of RNA secondary structure prediction by "X-Max-pMCC (N)" means the estimator of Eq. (23) with model X and number of samples N with respect to MCC. In the figure, we have also plotted the SEN-PPV curves of the γ-centroid estimator [10] with the CONTRAfold model ("CONTRAfold-gCentroid"; the black line) and with the McCaskill model ("McCaskill-gCentroid"; the gray line). The points and curve in gray and those in black indicate the McCaskill [16] and CONTRAfold [7] models, respectively. See Additional file 1, Figure S3 in the supplementary paper for the results of F-score.

Table 1 SEN, PPV, MCC and F-score for each prediction algorithm

	McCaskill model				CONTRAfold model			
γ	SEN	PPV	MCC	F-score	SEN	PPV	MCC	F-score
gc-pMCC	0.67	0.54	0.60	0.60	0.68	0.69	0.68	0.68
gc-pF	0.67	0.54	0.60	0.59	0.69	0.68	0.68	0.69
0.03125	0.34	0.78	0.52	0.48	0.09	0.90	0.28	0.16
0.0625	0.40	0.74	0.54	0.52	0.14	0.90	0.36	0.24
0.125	0.45	0.69	0.56	0.55	0.21	0.86	0.42	0.34
0.25	0.51	0.66	0.58	0.58	0.30	0.82	0.50	0.44
0.5	0.57	0.62	0.59	0.60	0.41	0.79	0.56	0.54
1.0	0.65	0.58	0.61	0.61	0.53	0.75	0.63	0.62
2.0	0.67	0.53	0.60	0.60	0.65	0.70	0.67	0.67
4.0	0.68	0.51	0.59	0.59	0.72	0.65	0.68	0.68
6.0	0.69	0.51	0.59	0.58	0.74	0.62	0.68	0.68
8.0	0.69	0.50	0.59	0.58	0.75	0.61	0.68	0.67
16.0	0.69	0.49	0.58	0.57	0.76	0.58	0.66	0.66
32.0	0.69	0.48	0.58	0.57	0.76	0.56	0.65	0.64
64.0	0.69	0.48	0.57	0.56	0.76	0.54	0.64	0.64
128.0	0.69	0.47	0.57	0.56	0.76	0.53	0.64	0.63
256.0	0.69	0.47	0.57	0.56	0.76	0.53	0.63	0.62
512.0	0.69	0.47	0.57	0.56	0.76	0.52	0.63	0.62
1024.0	0.69	0.47	0.57	0.56	0.76	0.52	0.63	0.62
RNAfold	0.65	0.50	0.57	0.57			-	
Simfold	0.64	0.51	0.57	0.57			-	
Sfold	0.65	0.58	0.62	0.62			-	

The rows labeled "gc-pMCC" and "gc-pF" indicate RNA secondary structure prediction with the γ-centroid estimator and the pseudo-expected MCC and F-score, respectively (Method M2). The rows below the dashed line indicate the results of the γ-centroid estimator [10] with various values of the γ parameter (given in the first column). Note that Simfold and Sfold employ a similar probabilistic model (based on energy parameters) for secondary structures to the McCaskill model.

Results of secondary structure prediction with γ- centroid estimator and pseudo-expected accuracy

Figure 3 shows the performance of RNA secondary structure prediction with the γ-centroid estimator and the pseudo-expected MCC/F-score (Method M2).

Table 2 Computational time in seconds

	Algorithm	McCaskill	CONTRAfold
1	γ-centroid with fixed γ	22	47
2	gCentroid-pMCC	36	59
3	Max-pMCC (1000)	178	303
4	Max-pMCC (10000)	1425	2391
5	Max-pMCC (100000)	13910	23291
6	Max-pMCC (1000000)	138987	232397

Total computational time in seconds for predicting secondary structures of all RNA sequences in the S-151Rfam dataset. The 1st row indicates the γ-centroid estimator [10] with a fixed γ parameter (1 for McCaskill model and 2 for CONTRAfold model). The 2nd row indicates the prediction of RNA secondary structure with the γ-centroid estimator and pseudo-expected MCC (Method M2). "Max-pMCC (N)" from the 3rd to 6th rows indicate the estimator of Eq. (23), that is, RNA secondary structure prediction by maximizing pseudo-expected MCC with stochastic sampling (Method M1) where N is the number of samples.

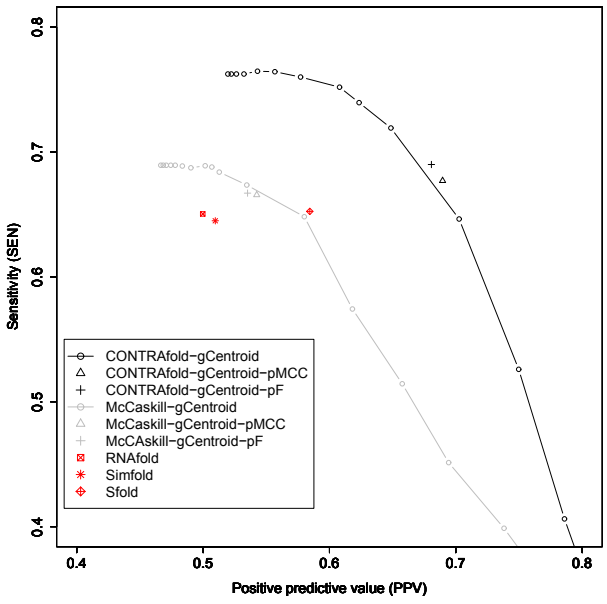

Figure 3 Performance of RNA secondary structure prediction with the γ-centroid estimator and the pseudo-expected MCC/F-score (Method M2). Performance of RNA secondary structure prediction with the γ-centroid estimator and the pseudo-expected MCC (F-score) (the estimator Eq. (24) with MCC (F-score); Method M2); "X-gCentroid-pMCC" ("X-gCentroid-pF") where X is the McCaskill or CONTRAfold model. The curves (X-gCentroid) indicate the performance of the γ-centroid estimator [10] with the McCaskill model and the CONTRAfold model. For comparison, we have also plotted the performance of RNAfold [14], Sfold [5] and Simfold [11] (red points). See Additional file 1, Figure S4 for performances of MEA estimators used in Do *et al.* [7].

When the McCaskill model is used, Method M2 is slightly worse than the γ-centroid estimator. However, the performance of Method M2 with the CONTRAfold model is slightly better than the performance of the γ-centroid estimator with the CONTRAfold model. (An example of both pre-dictions is shown in Additional file 1, Figure S5.)

It is also much better than the performance of RNAfold, Sfold and Simfold, all of which return a single prediction. Note that Method M2 with a fixed probabilistic model (e.g., the McCaskill model or the CONTRAfold model) generally achieves performance that differs from that of the γ-centroid estimator with the same model for any γ value. This is because Method M2 automatically selects the secondary structure with the best pseudo-expected MCC/F-score from a set of secondary structures given by the γ-centroid estimator for 17 γ values, while each point in a SEN-PPV curve of the γ- centroid estimator comes from a fixed γ-value.

Table 2 shows that the computational time of Method M2 is much shorter than for Method M1. This is because we do not need to perform any stochastic sampling in Method M2. In Figure 3, we also plotted the

performance of Sfold [5], Simfold [11] and RNAfold [14] (the points in red). The results indicate that the secondary structure predicted by Method M2 achieved better accuracy than those methods.

The comparison between the 2nd and 3rd rows in Table 2 indicates that there is only small overhead for the computation of the estimator of Method M2, compared with the γ-centroid estimator with a fixed γ parameter [10]. The reasons can be summarized as follows. The CYK-type algorithm of the Nussinov-style dynamic programming for computing a consistent RNA secondary structure is faster than the Inside-Outside-type algorithm for computing the base-pairing probability matrix in the γ-centroid estimator, even though both algorithms have the same computational complexity. Moreover, we do not need to employ the CYK-type algorithm for the γ- centroid estimator with $\gamma \leq 1$ because we only select the base-pairs whose base-pairing probability is larger than $1/(\gamma + 1)$ [10]. Also, the computation of the pseudo-expected MCC/F-score of a given secondary structure is fast enough when the base-pairing probability matrix is computed beforehand.

In summary, by combining the pseudo-expected accuracy with the γ-centroid estimator, we successfully predict the well-balanced secondary structure between SEN and PPV (with small overhead compared to CentroidFold) and the performance (with CONTRAfold model) is better than that of RNAfold, Simfold, Sfold and CentroidFold.

Discussion and Conclusion

In this study, we introduced the *pseudo*-expected accuracy, (with respect to commonly used accuracy measures in RNA secondary structure prediction: sensitivity, PPV, MCC or F-score) of a given RNA secondary structure under a probability distribution of possible secondary structures. The pseudo-expected accuracy can be computed much more easily than the expected accuracy, because it is computed using the base-pairing probability matrix of the RNA sequence. Although the pseudo-expected accuracy of a given secondary structure is not equal to the expected accuracy of the structure, our computational experiments have indicated that the pseudo-expected accuracy of a given secondary structure is a good approximation of the expected accuracy of the structure when SEN, PPV, MCC and F-score were used as the accuracy mea-sure. This finding is one of the contributions of this study, which has not been reported in previous research.

Based on this finding, we introduced the approximate estimator that maximizes the pseudo-expected accuracy of a prediction by stochastic sampling, which achieved favorable accuracy in our computational experiments. Although the computational cost of this estimator is much smaller than the estimator that maximizes the *expected* accuracy, it is still unacceptably slow. Therefore, we then proposed the combination of the pseudo-expected MCC/F-score and the γ-centroid estimator, which produces one well-balanced secondary structure with small computational overhead. The computational experiments indicate that this approach achieved the best accuracy among state-of-the-art tools. To employ the γ-centroid estimator in Method M2 is suitable because the γ-centroid estimator is able to represent a secondary structures with an arbitrary balance between the expected TP, TN, FP and FN by adjusting the parameter γ (see Eq. (18)). This, however, does not prove that there always exists a γ such that the γ-centroid estimator achieves the *best* pseudo-expected MCC or F-score. Note that the combination of the pseudo-expected MCC/F-score with the MEA-based estimator proposed by [7] is not suitable because the estimator has a *bias* to MCC and F-score, compared to the γ-centroid estimator [10].

Although the trade-ff between SEN and PPV is inherent, and MCC or F-score is not always the best choice of quality measure for predicted secondary structures, the proposed method (Method M2) can be applicable when only a single structure is required. The pseudo-expected MCC/F-score is also employed as a ranking measure of several predicted secondary structures.

Remarks about terminology: "maximum expected accuracy"

As we described in the Introduction section, the term "maximum (maximizing) expected accuracy" (MEA) has been used in a number of previous studies [6,7,10,26] as well as this study. From a mathematical viewpoint, the MEA (estimator) is a (point) estimator described as follows. Given a predictive space Y that contains all the possible candidate solutions of the target problem, a function $Acc(\theta, y)$ for $\theta \in Y$ and $y \in Y$, and a probability distribution $p(\theta|D)$ on Y given data D, then the estimator

$$\hat{\gamma} = \arg\max_{y \in Y} \int Acc(\theta, \gamma) p(\theta \mid D) d\theta \tag{25}$$

is introduced. When this estimator is called a "maximum expected accuracy" (MEA) estimator, $Acc(\theta, y)$ is equal to an accuracy measure (or is designed according to an accuracy measure) for a reference θ and a prediction y. This also implies that $p(\theta|D)$ is considered to be a probability distribution of *references*, which is misleading because $p(\theta|D)$ does not usually represent the distribution. In RNA secondary structure prediction, for example, The McCaskill model provides not a probability distribution of reference secondary structures but rather a full ensemble of possible secondary structures [16].

The estimator of Eq. (25) with a well-designed function $Acc(\theta, y)$ according to accuracy measures for a target problem and a probability distribution $p(\theta|D)$ of solutions empirically achieves better performance than other estimators such as the maximum likelihood estimator and the centroid estimator (i.e., the estimators that minimize the expected hamming difference) in RNA secondary structure predictions [7,10] and in alignments for biological sequences [25].

Difficulty of computing Eq. (20) with MCC and F-score

Eq. (20) with MCC and F-score can be rewritten as

$$\hat{y} = \arg\max_{\sigma \in S(x)} \frac{\sum_{i<j} p_{ij} I(\sigma_{ij} = 1)}{\sqrt{\sum_{i<j} I(\sigma_{ij} = 1)}} \text{ and} \quad (26)$$

$$\hat{y} = \arg\max_{\sigma \in S(x)} \frac{2 \times \sum_{i<j} p_{ij} I(\sigma_{ij} = 1)}{\sum_{i<j} I(\sigma_{ij} = 1) + \sum_{i<j} p_{ij}}, \quad (27)$$

respectively. Note that Eq. (26) is an *approximation* of Eq. (20) with MCC since TN (i.e., the number of true-negative base-pairs) is much larger than the others in RNA secondary structure predictions.

The denominators in both equations prevent division of this optimization problem into sub-problems, which is required to design a dynamic programming algorithm, and hence no efficient algorithms to compute Eqs. (26) and (27) have yet been devised. Note that the "argmax" operation for only the numerator can be efficiently solved by dynamic programming [33]. (This observation does not prove that there exists no efficient (polynomial time) algorithm for computing Eq. (20) with MCC and F-score.)

The proposed methods are extendable to other situations

We are able to introduce the pseudo-expected ac-curacy for *common* secondary structure prediction of multiple alignments of RNA sequences, because there are several probability distributions for the common secondary structures, for example, the RNAalifold model [34,35] and the Pfold model [36]. Also, the γ-centroid estimator can be extended to common secondary structure prediction [10], and the pseudo-expected MCC/F-score combined with the estimator is useful to predict the common secondary structure that balances between SEN and PPV (See [37]).

Recently, Lu *et al.* [6] proposed the relaxed SEN, PPV and MCC, where slippage of base-pair is al-lowed in computing those measures. It is possible to design the γ-centroid-type estimator that fits with those measures and also to introduce pseudo-expected accuracy of those measures. Moreover, the methods used in this paper can be extended to more general types of estimation problems (cf. [17]) with various accuracy measures that are defined by using TP, TN, FP and FN (cf. [29]).

The method presented in this paper can be applied to local alignments for biological sequences be-cause the γ-centroid estimator was also introduced in the problem [25]. In contrast to *global* alignment problems, the balance between SEN and PPV with respect to aligned bases is important in local alignment problems.

Additional material

Additional file 1: Supplementary Information for the main text. This file includes supplementary information for the main text.

Acknowledgements

This work was supported in part by the "Functional RNA Project" funded by the New Energy and Indus-trial Technology Development Organization (NEDO) of Japan and in part by a Grant-in-Aid for Scientific Research on Innovative Areas. The authors thank Drs/Profs H. Kiryu, M. C. Frith and T. Mituyama for valuable comments.

Author details

[1]Graduate School of Frontier Sciences, The University of Tokyo,5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan. [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST),2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan.

Authors' contributions

MH and KA conceived the study. MH developed the algorithms, performed the experiments and wrote the manuscript. KS implemented the algorithm in the CentroidFold software. All authors have read and approved the final manuscript.

Received: 2 July 2010 Accepted: 30 November 2010
Published: 30 November 2010

References

1. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, , **33 Database:** 121-124.
2. Andronescu M, Bereg V, Hoos H, Condon A: **RNA STRAND: the RNA secondary structure and statistical analysis database.** *BMC Bioinformatics* 2008, **9**:340.
3. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A: **Rfam: updates to the RNA families database.** *Nucleic Acids Res* 2009, **37**: D136-140.
4. Schroeder SJ: **Advances in RNA structure prediction from sequence: new tools for generating hypotheses about viral RNA structure-function relationships.** *J Virol* 2009, **83**:6326-6334.
5. Ding Y, Chan CY, Lawrence CE: **Sfold web server for statistical folding and rational design of nucleic acids.** *Nucleic Acids Res* 2004, , **32 Web Server:** 135-141.
6. Lu ZJ, Gloor JW, Mathews DH: **Improved RNA secondary structure prediction by maximizing expected pair accuracy.** *RNA* 2009, **15**:1805-1813.
7. Do C, Woods D, Batzoglou S: **CONTRAfold: RNA secondary structure prediction without physics-based models.** *Bioinformatics* 2006, **22**:e90-98.
8. Engelen S, Tahi F: **Tfold: efficient in silico prediction of non-coding RNA secondary structures.** *Nucleic Acids Res* 2010, **38**:2453-2466.

9. Parisien M, Major F: **The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data.** *Nature* 2008, **452**:51-55.
10. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K: **Prediction of RNA secondary structure using generalized centroid estimators.** *Bioinformatics* 2009, **25**:465-473.
11. Andronescu M, Condon A, Hoos H, Mathews D, Murphy K: **Efficient parameter estimation for RNA secondary structure prediction.** *Bioinformatics* 2007, **23**:19-28.
12. Zuker M, Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.** *Nucleic Acids Res* 1981, **9**:133-148.
13. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31(13)**:3406-3415.
14. Hofacker I, Fontana W, Stadler P, Bonhoeffer S, Tacker M, Schuster P: **Fast folding and comparison of RNA secondary structures.** *Monatsh Chem* 1994, **125**:167-188.
15. Mathews D, Disney M, Childs J, Schroeder S, Zuker M, Turner D: **Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.** *Proc Natl Acad Sci USA* 2004, **101**:7287-7292.
16. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29(6-7)**:1105-1119.
17. Carvalho L, Lawrence C: **Centroid estimation in discrete high-dimensional spaces with applications in biology.** *Proc Natl Acad Sci USA* 2008, **105**:3209-3214.
18. Ding Y, Chan C, Lawrence C: **RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble.** *RNA* 2005, **11**:1157-1166.
19. Hamada M, Sato K, Kiryu H, Mituyama T, Asai K: **Pre-dictions of RNA secondary structure by combining homologous sequence information.** *Bioinformatics* 2009, **25**:i330-338.
20. Kiryu H, Kin T, Asai K: **Robust prediction of consensus secondary structures using averaged base pairing probability matrices.** *Bioinformatics* 2007, **23**:434-441.
21. Seemann S, Gorodkin J, Backofen R: **Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments.** *Nucleic Acids Res* 2008, **36**:6355-6362.
22. Hamada M, Sato K, Kiryu H, Mituyama T, Asai K: **CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score.** *Bioinformatics* 2009, **25**:3236-3243.
23. Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L: **Fast statistical alignment.** *PLoS Comput Biol* 2009, **5**:e1000392.
24. Bradley RK, Pachter L, Holmes I: **Specific alignment of structured RNA: stochastic grammars and sequence annealing.** *Bioinformatics* 2008, **24**:2677-2683.
25. Frith MC, Hamada M, Horton P: **Parameters for Accurate Genome Alignment.** *BMC Bioinformatics* 2010, **11**:80.
26. Kall L, Krogh A, Sonnhammer EL: **An HMM posterior decoder for sequence feature prediction that includes homology information.** *Bioinformatics* 2005, **21(Suppl 1)**:i251-257.
27. Michal N, Tomas V, Brona B: **The Highest Expected Reward Decoding for HMMs with Application to Recombination Detection.** *arXiv.org* 2010 [http://arxiv.org/abs/1001.4499].
28. Gross S, Do C, Sirota M, Batzoglou S: **CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction.** *Genome Biol* 2007, **8**:R269.
29. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview.** *Bioinformatics* 2000, **16**:412-424.
30. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological sequence analysis* Cambridge, UK: Cambridge University press; 1998.
31. Ding Y, Lawrence CE: **A statistical sampling algorithm for RNA secondary structure prediction.** *Nucleic Acids Res* 2003, **31**:7280-7301.
32. Sato K, Hamada M, Asai K, Mituyama T: **CENTROID- FOLD: a web server for RNA secondary structure prediction.** *Nucleic Acids Res* 2009, **37**:W277-280.
33. Holmes I, Durbin R: **Dynamic programming alignment accuracy.** *J Comput Biol* 1998, **5**:493-504.
34. Bernhart S, Hofacker I, Will S, Gruber A, Stadler P: **RNAalifold: improved consensus structure pre-diction for RNA alignments.** *BMC Bioinformatics* 2008, **9**:474.
35. Hofacker IL, Fekete M, Stadler PF: **Secondary structure prediction for aligned RNA sequences.** *J Mol Biol* 2002, **319(5)**:1059-1066.
36. Knudsen B, Hein J: **Pfold: RNA secondary structure prediction using stochastic context-free grammars.** *Nucleic Acids Res* 2003, **31(13)**:3423-3428.
37. Hamada M, Sato K, Asai K: **Improving the ac-curacy of predicting secondary structure for aligned RNA sequences.** *Nucleic Acids Res* 2010.

doi:10.1186/1471-2105-11-586
Cite this article as: Hamada *et al.*: Prediction of RNA secondary structure by maximizing pseudo-expected accuracy. *BMC Bioinformatics* 2010 **11**:586.