Direct Updating of an RNA Base-Pairing Probability Matrix with Marginal Probability Constraints

MICHIAKI HAMADA

ABSTRACT

A base-pairing probability matrix (BPPM) stores the probabilities for every possible base pair in an RNA sequence and has been used in many algorithms in RNA informatics (e.g., RNA secondary structure prediction and motif search). In this study, we propose a novel algorithm to perform iterative updates of a given BPPM, satisfying marginal probability constraints that are (approximately) given by recently developed biochemical experiments, such as SHAPE, PAR, and FragSeq. The method is easily implemented and is applicable to common models for RNA secondary structures, such as energy-based or machine-learning-based models. In this article, we focus mainly on the details of the algorithms, although preliminary computational experiments will also be presented.

Key words: algorithms, alignment, RNA, secondary structure, sequence analysis.

1. INTRODUCTION

NON-CODING RNAs (e.g., miRNA, snoRNA, piRNA, and lincRNA) play many important roles in life (Gardner et al., 2011), some of which are related to disease (Esteller, 2011). The secondary and tertiary structures of RNA sequences are a key to understanding their functions, and much research has been conducted in this field (Leontis and Westhof, 2012).

The *base-pairing probability matrix* (BPPM) of an RNA sequence, which stores all the probabilities for possible base pairs in an RNA sequence (see Section 2 for the detailed definition), has played an essential role in a number of algorithms in RNA informatics, including RNA (common) secondary structure predictions (Seemann et al., 2008,) Lu et al., 2009; Hamada et al., 2009b, 2011c; Sato et al., 2011, multiple alignment of RNA sequences (Hofacker et al., 2004; Katoh and Toh, 2008; Hamada et al., 2009a; Sahraeian and Yoon, 2011), RNA–RNA interaction predictions (Kato et al., 2010; Seemann et al., 2011), RNA motif search (Hamada et al., 2006), and miRNA gene finding (Terai et al., 2007) (Wei et al., 2011). Estimating accurate base-pairing probabilities, therefore, has the potential to improve those algorithms without modifying them.

On the other hand, the recent advent of experimental techniques to probe RNA structure by high-throughput sequencing—SHAPE (Low and Weeks, 2010); PARS (Kertesz et al., 2010); FragSeq (Underwood et al., 2010)—has enabled genome-wide measurements of RNA structure, leading to the world of the “RNA structurome” (Wan et al., 2011). Those experimental techniques (SHAPE, PARS, and FragSeq)

The University of Tokyo, Graduate School of Frontier Science, Kashiwa, Japan; and National Institute of Advanced Industrial Science and Technology (AIST), Computational Biology Research Center, Tokyo, Japan.

stochastically estimate the flexibility of an RNA strand, which can be considered as a kind of *loop probability* for every nucleotide in an RNA sequence. Remarkably, secondary structures of long RNA sequences—HIV-1 (Watts et al., 2009), HCV (hepatitis C virus) (Pang et al., 2011), and lincRNA (the steroid receptor RNA activator [SRA]) (Novikova et al., 2012)—were recently determined by combining those experimental techniques with computational approaches. It is therefore important to incorporate this experimental information into RNA informatics, especially structural analysis for RNAs. Washietl et al. (2012) have proposed a method to modify the Turner energy parameters to be consistent with the experimental (SHAPE) data. Deigan et al. (2009) incorporated experimental information (given by SHAPE) into energy calculations of RNA secondary structure. (See Low and Weeks [2010] for recent reviews.) Quarrier et al. (2010) proposed a method to generate “decoy” RNA structures by stochastic sampling and evaluating whether the structure fits the constraints. However, there are many models for RNA secondary structures, such as Turner’s energy model (McCaskill, 1990; Hofacker et al., 1994; Mathews et al., 1999) and the (machine learning) stochastic model (Do et al., 2006; Andronescu et al., 2007, 2010), and we must implement the algorithms separately for each model.

Against this background, we propose a novel lightweight and simple algorithm for updating the base-pairing probability matrix according to marginal probability constraints after formulating the problem mathematically. We found that the problem has a similar structure to estimating an input–output (IO) table in macroeconomics, which is called the matrix balancing problem (Censor and Zenios, 1997; Morioka and Tsuda, 2011). Therefore, our algorithms are based on the RAS algorithms used in macroeconomics. The algorithm is not only simple and fast but also easily implemented and widely applicable to models of RNA secondary structures that produce a base-pairing probability matrix. Most tools for predicting RNA secondary structures have an option to produce a base-pairing probability matrix.¹

This article is organized as follows. In Section 2, we mathematically formulate the targeted problem, which is followed by algorithms to solve the problem in Section 3, including the proof of validity of the algorithms. Preliminary computational experiments for confirming the effectiveness of the algorithms are discussed in Section 4.

2. PROBLEM FORMULATION

The following representation of RNA secondary structures is often used (see, e.g., Hamada et al., 2011b).

Definition 1 (The space of secondary structures of an RNA sequence x : $\mathcal{S}(x)$). For an RNA sequence x , a secondary structure of x is represented as an upper triangular binary-valued matrix, $y = \{y_{ij}\}_{1 \leq i < j \leq |x|}$, where $y_{ij} = 1$ means x_i and x_j (the i th and j th bases of x) form a base pair and $y_{ij} = 0$ means x_i and x_j do not form a base pair. $\mathcal{S}(x)$ denotes the space of possible secondary structures of x .

A probability distribution, $p(\theta|x)$, on the space $\mathcal{S}(x)$ can be introduced by using one of the following probabilistic models: McCaskill model (McCaskill, 1990), based on Turner’s energy model of RNA secondary structures (Mathews et al., 1999); CONTRAfold model (Do et al., 2006), based on conditional random fields, (CRFs); Boltzmann likelihood (BL) model (Andronescu et al., 2007, 2010); and others (Sato et al., 2009; Zakov et al., 2011; Rivas et al., 2012).

Definition 2 (Base-pairing probability matrix, BPPM). For an RNA sequence x and a probability distribution $p(\theta|x)$ on $\mathcal{S}(x)$, an upper triangular matrix $M = \{p_{ij}\}_{1 \leq i < j \leq |x|}$, where $p_{ij} = \sum_{\theta \in \mathcal{S}(x)} I(\theta_{ij} = 1)p(\theta|x)$, is called a base-pairing probability matrix (BPPM) of x . Here, $I(\cdot)$ is the indicator function that takes a value of 1 or 0, depending on whether the condition constituting its argument is true or false.

The base-pairing probability matrix includes the information about the ensemble of secondary structures of a given RNA sequence, while a predicted secondary structure [e.g., a minimum free energy (MFE)

¹For example, in the case of RNAfold (Hofacker et al., 1994), we can compute a BPPM by using RNAfold with the `-p` option; for CentroidFold (Hamada et al., 2009b), we use `centroid_fold` with the `-posteriors` option.

structure] does not. This ensemble information is useful for developing algorithms for RNA informatics, because every secondary structure (even the MFE structure) has extremely low probability (Carvalho and Lawrence, 2008; Hamada et al., 2011b). It is well known that the base-pairing probability matrix for a given RNA sequence can be computed efficiently (by employing a dynamic programming [DP] algorithm) for the previously mentioned probability distributions on $\mathcal{S}(x)$ (see McCaskill [1990] for the details).

Next, we introduce a kind of distance between two base-pairing probability matrices as follows.

Definition 3 (Kullback-Leibler [KL] distance between two BPPMs). For two base-pairing probability matrices, $A = \{a_{ij}\}_{i < j}$ and $B = \{b_{ij}\}_{i < j}$,

$$D(A, B) = \sum_{i < j} [a_{ij} \log a_{ij} - a_{ij} \log b_{ij} - a_{ij} + b_{ij}] \quad (1)$$

is called the Kullback-Leibler (KL) distance between A and B .

This distance was recently used in the study by Janssen et al. (2011).

For use in the following discussion, we define

$$\mathcal{I}_i (= \mathcal{I}_i^{(x)}) := \{(i, j) | 1 \leq i < j \leq |x|\} \cup \{(j, i) | 1 \leq j < i \leq |x|\} \quad (2)$$

for an RNA sequence x and a position $i \in [1, |x|]$ in x . In other words, $\mathcal{I}_i$ is the set of base pairs where the position of one base in the pair is equal to i .

Definition 4 (Marginal probability constraints for BPPM). A vector $\mathbf{q} = \{q_i\}_{1 \leq i \leq |x|}$ ($q_i \in [0, 1]$) is called a marginal probability constraint vector when q_i is expected to be equal (or similar) to $\sum_{(k, l) \in \mathcal{I}_i} p_{kl}$ for a base-pairing probability matrix $P = \{p_{ij}\}_{i < j}$.

$1 - q_i$ is taken as the *loop probability* of the i th position in the RNA sequence x , which indicates the sum of probabilities of secondary structures whose i th base is a loop (in other words, is not in a base pair). Correspondingly, q_i is the probability of a base pair at the i th position. Experimental techniques, such as SHAPE (Low and Weeks, 2010), give the strand flexibility of every nucleotide in an RNA sequence and, therefore, approximately give the loop probabilities.

Now our problem is formulated as follows. We would like to find a new BPPM satisfying the marginal constraints that is nearest to the initial BPPM.

Problem 1. Given an initial base-pairing probability matrix A (Definition 2) of an RNA sequence x and a constraint vector $\mathbf{q} = \{q_i\}_{1 \leq i \leq |x|}$ (Definition 4), solve the optimization problem

$$\begin{aligned} \hat{X} = \{x_{ij}\} &= \underset{X}{\operatorname{argmin}} D(X, A) \\ \text{subject to } \sum_{(k, l) \in \mathcal{I}_i} x_{kl} &= q_i \quad \text{for all } i = 1, \dots, |x| \\ x_{ij} &\geq 0 \quad \text{for all } i < j \end{aligned}$$

In practice, the initial matrix A is found computationally, and the constraints $\mathbf{q}$ are given by biological experiments.

This problem is related to the estimation of an input–output (IO) table in macroeconomics. In economics, an input–output model is a quantitative economic technique that represents the interdependencies between different branches of the national economy or between branches of different, even competing, economies. The matrix-balancing problem is the problem of updating an input–output table subject to given sums of every row and column, and this has a similar structure to Problem 1.

In contrast to the matrix-balancing problem, biological experiments such as SHAPE provide constraints that have a large amount of noise and are, therefore, not so precise. Problem 1, therefore, seems too strict because it tries to obtain a base-pairing probability matrix that is *completely* consistent with the constraints. We, therefore, relax the problem as follows.

Problem 2. Fix $\varepsilon \in [0, 1]$. Given an initial base-pairing probability matrix $A = (p_{ij})_{1 \leq i < j \leq |x|}$ of an RNA sequence x and a constraint vector $\mathbf{q} = \{q_i\}_{1 \leq i \leq |x|}$, solve the optimization problem

$$\hat{X} = \{x_{ij}\} = \underset{X}{\operatorname{argmin}} D(X, A) \quad (3)$$

$$\text{subject to } \underline{u}_i^{(\varepsilon)} \leq \sum_{(k,l) \in \mathcal{I}_i} x_{kl} \leq \bar{u}_i^{(\varepsilon)} \quad \text{for all } i = 1, \dots, |x| \quad (4)$$

$$p_{ij} \geq 0 \quad \text{for all } i, j \quad (5)$$

where

$$\underline{u}_i^{(\varepsilon)} = \max(0, q_i - \varepsilon) \quad \text{and} \quad \bar{u}_i^{(\varepsilon)} = \min(1, q_i + \varepsilon)$$

Clearly, Problem 2, with $\varepsilon = 0$, is equivalent to Problem 1. Note that, in Problem 2, the lower and upper bounds with respect to marginal probability constraints can be chosen arbitrarily. Therefore, if there are missing data in the marginal constraint vector (Definition 4), we can employ a version of Problem 2 in which $\underline{u}_i^{(\varepsilon)} = 0$ and $\bar{u}_i^{(\varepsilon)} = 1$ for each position i that corresponds to missing data. This means that there is no constraint for the i th position.

3. ALGORITHMS

In this section, we present algorithms for solving Problem 1 and Problem 2, both of which are simple iterative algorithms. In Algorithm 1, every base-pairing probability whose index is in $\mathcal{I}_i$ is iteratively re-scaled in order to satisfy the i th marginal condition (line 19). Note that Algorithm 1 is similar to the RAS algorithm in macroeconomics (see, e.g., Morioka and Tsuda, 2011). On the other hand, in Algorithm 2, the median value is used to re-scale the base-pairing probability (line 22). It is easily seen that Algorithm 1 is a special case of Algorithm 2. In practice, we must avoid having $\text{total} = 0$ and $q[i] \neq 0$ in line 17 of Algorithm 1, so a small probability (e.g., 0.0001) is added to every (A-U, G-C, G-U) base pair of an initial BPPM. The following theorem assures the validity of these algorithms.

Algorithm 1: Iterative solution for Problem 1

```

1: procedure ITERATION( $A, \mathbf{q}, \text{maxitr}$ )                                     ▷  $A$ : initial BPPM;  $\mathbf{q}$ : constraint vector
2:    $X \leftarrow A$ 
3:   for  $v = 1$  to  $\text{maxitr}$  do
4:      $X' \leftarrow X$ 
5:     UPDATE( $X, \mathbf{q}$ )
6:     if  $\text{KL}(X, X') < \varepsilon$  then return  $X$                                ▷ Compute the KL distance (see Eq.(1))
7:   end if
8: end for
9: return  $X$ 
10: end procedure
11: procedure UPDATE( $X, \mathbf{q}$ )                                              ▷ Obtain an updated BPPM from a previous BPPM
12:   for  $i = 1$  to  $L$  do                                                ▷  $L$  is the length of the RNA sequence
13:      $\text{total} \leftarrow 0$ 
14:     for all  $(k, l) \in \mathcal{I}_i$  do                                         ▷ See Eq.(2) for the definition of  $\mathcal{I}_i$ 
15:        $\text{total} \leftarrow \text{total} + X[k][l]$ 
16:     end for
17:      $C \leftarrow q[i] / \text{total}$                                           ▷ Re-scaling factor
18:     for all  $(k, l) \in \mathcal{I}_i$  do
19:        $X[k][l] \leftarrow X[k][l] * C$                                    ▷ Re-scale the value in  $\mathcal{I}_i$ 
20:     end for
21:   end for
22: end procedure

```

Theorem 1. *Algorithm 1 and Algorithm 2 converge to the solutions of Problem 1 and Problem 2, respectively.*

Proof. Because Algorithm 1 is a special case of Algorithm 2, we will prove this theorem for Algorithm 2 only.

Algorithm 2: Iterative solution for Problem 2

```

1: procedure ITERATION( $A, \mathbf{q}, \varepsilon, \text{maxitr}$ )                                 $\triangleright$   $A$ : initial BPPM;  $\mathbf{q}$ : constraint vector;  $\varepsilon$ : relaxing parameter
2:    $X \leftarrow A$ 
3:   for  $i = 1$  to  $L$  do                                                 $\triangleright$   $L$  is the length of the RNA sequence
4:      $r[i] \leftarrow 1$                                                $\triangleright$  Initialize a vector  $\mathbf{r}$ 
5:   end for
6:   for  $v = 1$  to  $\text{maxitr}$  do
7:      $X' \leftarrow X$ 
8:     UPDATE( $X, \underline{u}^{(e)}, \overline{u}^{(e)}, \mathbf{r}$ )
9:     if  $\text{KL}(X, X') < \varepsilon$  then return  $X$                              $\triangleright$  Compute the KL distance (see Eq.(1))
10:    end if
11:  end for
12:  return  $X$ 
13: end procedure
14: procedure UPDATE( $X, \underline{u}, \overline{u}, \mathbf{r}$ )                                     $\triangleright$  Updating BPPM
15:   for  $i = 1$  to  $L$  do                                                 $\triangleright$   $L$  is the length of the RNA sequence
16:      $\text{total} \leftarrow 0$ 
17:     for all  $(k, l) \in \mathcal{I}_i$  do                                        $\triangleright$  See Eq.(2) for the definition of  $\mathcal{I}_i$ 
18:        $\text{total} \leftarrow \text{total} + X[k][l]$ 
19:     end for
20:      $\underline{u}[i] \leftarrow \underline{u}[i]/\text{total}$ 
21:      $\overline{u}[i] \leftarrow \overline{u}[i]/\text{total}$ 
22:      $C \leftarrow \text{median}(\underline{u}[i], \overline{u}[i], r[i])$                          $\triangleright$  Median of three values becomes the re-scaling factor
23:     for all  $(k, l) \in \mathcal{I}_i$  do
24:        $X[k][l] \leftarrow X[k][l] * C$                                  $\triangleright$  Re-scale the value in  $\mathcal{I}_i$ 
25:     end for
26:      $r[i] \leftarrow r[i]/C$ 
27:   end for
28: end procedure

```

First, we vectorize the base-pairing probability matrix $X = \{x_{ij}\}_{1 \leq i < j \leq n}$ as $x = \{x_s\}_{s=1}^{n(n-1)/2}$ where x_{ij} ($i < j$) corresponds to x_s with $s = (2n-i)(i-1)/2 + (j-i)$. $A = \{a_{ij}\}_{1 \leq i < j \leq n}$ is also vectorized to $a = \{a_s\}_{s=1}^{n(n-1)/2}$. Second, we introduce the following $n \times n(n-1)/2$ binary matrix.

$$\Phi = \left(\begin{array}{c|c|c|c|c|c} \overbrace{\begin{matrix} 1 & 1 & 1 & \cdots & 1 \end{matrix}}^{n-1} & \overbrace{\begin{matrix} 1 & 1 & \cdots & 1 \end{matrix}}^{n-2} & \overbrace{\begin{matrix} 1 & \cdots & 1 \end{matrix}}^{n-3} & \cdots & \overbrace{\begin{matrix} 1 & 1 \end{matrix}}^2 & \overbrace{\begin{matrix} 1 \end{matrix}}^1 \\ \hline 1 & & & & & \\ \hline & 1 & & & & \\ \hline & & 1 & & & \\ \hline & & & 1 & & \\ \hline & & & & 1 & 1 \\ \hline & & & & & 1 & 1 \end{array} \right) \quad \left. \vphantom{\begin{matrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & 1 \\ & & & & & 1 & 1 \end{matrix}} \right\} n$$

where each blank in the matrix is equal to 0. It is easily seen that the i th element of Φx is equal to $\sum_{(k,l) \in \mathcal{I}_i} x_{kl}$. Problem 2 can, therefore, be recast as solving the optimization problem

$$\begin{aligned} \hat{x} &= \underset{x}{\operatorname{argmin}} f(x; a) \\ \text{where } f(x; a) &:= D(x, a) = \sum_{s=1}^{n(n-1)/2} (x_s \log x_s - x_s \log a_s - x_s) \\ \text{subject to } \underline{u}^{(e)} &\leq \Phi x \leq \bar{u}^{(e)} \text{ and } x \geq 0 \end{aligned}$$

where $\underline{u}^{(e)} = (\underline{u}_i^{(e)})$ and $\bar{u}^{(e)} = (\bar{u}_i^{(e)})$. Note that f is a kind of Bregman function (Bregman, 1967), and there is a unique generalized projection onto a hyperplane. By using Theorem 6.4.1 in Censor and Zenios (1997), the following iteration converges to the solution of Problem 2.

Initialization step: Initialize x^0 and z^0 as follows.

$$\begin{aligned} x^0 &\in U := \{x | x \geq 0, \exists z \in \mathbb{R}^n \text{ s.t. } \nabla f(x) = -\Phi^T z\} \\ \nabla f(x^0) &= -\Phi^T z^0 \end{aligned}$$

Iteration step: For $\nu = 0, 1, 2, \dots$,

$$\nabla f(x^{\nu+1}) = \nabla f(x^\nu) + c_\nu \phi^{i(\nu)} \quad (6)$$

$$z^{\nu+1} = z^\nu - c_\nu e^{i(\nu)} \quad (7)$$

$$c_\nu = \min(z_{i(\nu)}^\nu, \delta_\nu, \gamma_\nu) \quad (8)$$

where $i(\nu) = (\nu \bmod n + 1)$, ϕ^i is the i th row vector of Φ and e^i is the i th unit vector in $\mathbb{R}^n$ (i.e., the i th element of e is 1; 0 otherwise). In Equation (8), δ_ν and γ_ν are Lagrangian multipliers for the optimization problems

$$\text{Minimize } f(x; x^\nu) \text{ subject to } \sum_k \Phi_{i(\nu)k} x_k = \underline{u}_{i(\nu)}^{(e)} \quad (9)$$

and

$$\text{Minimize } f(x; x^\nu) \text{ subject to } \sum_k \Phi_{i(\nu)k} x_k = \bar{u}_{i(\nu)}^{(e)} \quad (10)$$

respectively, where Φ_{ik} is the (i,k) -element in the matrix Φ .

We will now prove that the steps given above correspond to Algorithm 2. In the initialization step, we set $z^0 = 0$ and hence $x^0 = a$. The Lagrangian for Equation (9) is defined as

$$L(x, \gamma) = \sum_s x_s (\log x_s - \log x_s^\nu - 1) - \delta_\nu \left(\sum_s \Phi_{i(\nu)s} x_s - \underline{u}_{i(\nu)}^{(e)} \right). \quad (11)$$

According to the KKT conditions (Kuhn and Tucker, 1951), the equations

$$\frac{\partial L}{\partial x_s} = 0 \text{ (for all } s) \text{ and } \frac{\partial L}{\partial \delta_\nu} = 0$$

are satisfied. Then,

$$\frac{\partial L}{\partial x_s} = \log x_s - \log x_s^\nu - \delta_\nu \Phi_{i(\nu)s} = 0,$$

leads to

$$x_s = x_s^\nu \exp(\delta_\nu \Phi_{i(\nu)s}). \quad (12)$$

By substituting Equation (12) into Equation (11), we obtain

$$L = - \sum_s x_s^\nu \exp(\delta_\nu \Phi_{i(\nu)s}) + \delta_\nu \underline{u}_{i(\nu)}^{(e)}. \quad (13)$$

Differentiating Equation (13) with respect to δ_ν and setting the result equal to 0 gives

$$\frac{\partial L}{\partial \delta_\nu} = \sum_s \Phi_{i(\nu)s} x_s^\nu \exp(\delta_\nu \Phi_{i(\nu)s}) - \underline{u}_i^{(\varepsilon)} = 0$$

which leads to

$$\exp(\delta_\nu) = \frac{\underline{u}_i^{(\varepsilon)}}{\sum_{s: \Phi_{i(\nu)s} = 1} x_s^\nu}.$$

Similarly, we obtain

$$\exp(\gamma_\nu) = \frac{\underline{u}_i^{(\varepsilon)}}{\sum_{s: \Phi_{i(\nu)s} = 1} x_s^\nu}.$$

By taking exponentials of Equations (7) and (8), we have

$$\exp(z_{i(\nu)}^{\nu+1}) = \frac{\exp(z_{i(\nu)}^\nu)}{\exp(c_\nu)} \quad (14)$$

$$\exp(c_\nu) = \text{median}(\exp(z_{i(\nu)}^\nu), \exp(\delta_\nu), \exp(\gamma_\nu)) \quad (15)$$

respectively. Equations (14) and (15) lead to lines 26 and 22 in Algorithm 2, respectively. Finally,

$$x_s^{\nu+1} = x_s^\nu \exp(c_\nu \phi_s^{i(\nu)}) = x_s^\nu \exp(c_\nu) \exp(\phi_s^{i(\nu)}) \quad (16)$$

leads to line 24 in Algorithm 2, and the proof is finished. $\blacksquare$

4. COMPUTATIONAL EXPERIMENTS

We have implemented a Perl script for updating a given initial base-pairing probability matrix in order to satisfy marginal constraints (Algorithms 1 and 2). For the updated base-pairing probability matrix, the γ -centroid estimator (Hamada et al., 2009b) was employed to predict a secondary structure from the BPPM. More specifically, we maximize the sum of the (updated) base-pairing probabilities for base pairs in the predicted secondary structure, in which base pairs whose probability is larger than $1/(\gamma + 1)$ are allowed to form a base pair (Hamada et al., 2009b). The γ is a parameter used to adjust the balance between sensitivity and positive predictive value of predicted base-pairs (a smaller γ produces fewer base pairs). CentroidFold software (<http://ncrna.org/software/centroidfold/download/>), which is an implementation of the γ -centroid estimator, was utilized. In the following experiments, the initial BPPM (A in Algorithms 1 and 2) is given by computational methods—either the Boltzmann likelihood (BL) model (Andronescu et al., 2010) or the CONTRAfold model (Do et al., 2006), both of which are implemented in the CentroidFold software.

To confirm that our algorithm works correctly, we first conducted computational experiments with *perfect* constraints. Here, the constraints $\mathbf{q}$ (in Definition 4) are given according to the reference secondary structure: $q_i = 1$ when the base at position i forms a base pair in the reference secondary structure and $q_i = 0$ otherwise. As a benchmark dataset, we used the “S151 Rfam dataset,” which has been employed in many studies (Do et al., 2006; Hamada et al., 2011a).

The results of the first experiment are shown in Figure 1. As performance measures, we used the sensitivity (SEN) and positive predictive value (PPV): $\text{SEN} = \text{TP}/(\text{TP} + \text{FN})$ and $\text{PPV} = \text{TP}/(\text{TP} + \text{FP})$ where TP, FP, and FN are the numbers of true-positive base pairs, false-positive base pairs, and false-negative base pairs, respectively. We computed those values for each predicted secondary structure only once for the entire dataset. In the γ -centroid estimator, 17 γ values, $\gamma \in \{2^k : -5 \leq k \leq 10, k \in \mathbb{Z}\} \cup \{6\}$, were used to draw sensitivity-PPV curves. For each probabilistic model for RNA secondary structures, when ε is close to 1, the performance approaches the case in which secondary structure is predicted only from the initial base-pairing probability matrix (i.e., the performance of the original CentroidFold software), because the initial BPPM itself satisfies the marginal constraint (Eq. 4) when ε is large. Examples of updated base-pairing probability matrices for every ε value are shown in Figure 2; tRNA sequences were utilized in the experiments. Table 1

FIG. 1. Benchmark experiments for *perfect* constraints for various ε values (the relaxation parameter in Problem 2). The γ -centroid estimators (Hamada et al., 2009b) are used when estimating secondary structures from the updated base-pairing probability matrix. We utilized the S151 Rfam dataset in this experiment, and constraints are given according to reference secondary structures. Initial base-pairing probability matrices are computed using (a) the BL model (Andronescu et al., 2010) and (b) the CONTRAfold model (Do et al., 2006).

shows that the computational time is shorter for larger ε , because the number of iterations is smaller for larger epsilon. On the other hand, the computational time for $\varepsilon = 0$ is slightly faster than the case of $\varepsilon = 0.01$, because, in the former case, the simpler Algorithm 1 is used rather than Algorithm 2.

Second, we conducted experiments using marginal constraints with missing data, where missing data is simulated in the perfect constraints described above. The ratio of missing positions varies from 0% to 100% in 10% steps. Figure 3 shows the results, indicating that the performance gradually approaches that of the original CentroidFold as the proportion of missing data increases.

Third, we conducted experiments for marginal probability constraints with *noise*. In this dataset, we simulated uniform noise of 10% for positions in loop regions and 30% noise in base-pair regions, with respect to perfect constraints. Figure 4 shows the results. This indicates that the performance for $\varepsilon = 0$ was worse than that for $\varepsilon \neq 0$ in a part of sensitivity-PPV curve, which indicates that a non-zero ε value could have the potential to improve the prediction accuracy when constraints include noise.

TABLE 1. COMPUTATIONAL TIME AND ITERATION NUMBER

ε	0	0.01	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99	0.999	Orig
<i>BL model</i>															
Time	594	609	389	266	212	185	168	157	136	114	93	74	67	67	15
Iter	109.8	87.6	56.2	38.6	26.3	20.4	16.6	12.9	9.6	6.4	3.6	1.9	1.2	1.1	NA
<i>CONTRAfold model</i>															
Time	496	508	466	273	213	200	192	189	177	157	134	131	122	118	30
Iter	102.5	71.0	55.3	35.0	23.7	19.2	15.6	11.9	8.7	5.5	3.0	1.6	1.1	1.0	NA

The row “Time” contains the total computational times in seconds for all secondary structures in the S151 Rfam dataset (Do et al., 2006) for 17 values of γ , and the row “Iter” contains the average number of iterations in the proposed algorithms. The column “Orig” is the performance of the original CentroidFold (Hamada et al., 2009b) with the specified model. In this experiment, we utilized the perfect constraints, which were also used to produce Figure 1. We employed two probabilistic models for RNA secondary structures: the CONTRAfold model (Do et al., 2006) and the Boltzmann likelihood (BL) model (Andronescu et al., 2010). e is set to 10^{-4} in Algorithms 1 and 2. We used a computer with a 2.67 GHz Intel Xeon Processor X5550, 24 GB of RAM, and a Linux operating system.

FIG. 2. Updated base-pairing probability matrices of a tRNA sequence (RF00005) for $\varepsilon = 0, 0.2, 0.4, 0.6, 0.8$. Perfect constraints were utilized in this example. The bottom and rightmost figure (labeled “Initial”) is the initial base-pairing probability matrix (i.e., A in Algorithms 1 and 2).

FIG. 3. Benchmark experiments for constraints including missing data, which are simulated by inserting missing values in the perfect constraints. The ratio of missing positions varies 0% to 100% in 10% steps. The γ -centroid estimators (Hamada et al., 2009b) are used when estimating secondary structures from the updated base-pairing probability matrix. We utilized the S151 Rfam dataset in this experiment and, constraints are given according to reference secondary structures. Initial base-pairing probability matrices are computed using (a) the BL model and (b) the CONTRAfold model.

FIG. 4. Benchmark experiments for constraints with noise using various values of ε (the relaxation parameter). We simulated uniform noise of 10% for positions in loop regions and 30% noise in base-pair regions, with respect to perfect constraints. The γ -centroid estimators (Hamada et al., 2009b) are used when estimating secondary structures from updated base-pairing probability. We utilized the S151 Rfam dataset in this experiment, and constraints are given according to reference secondary structures. Initial base-pairing probability matrices are computed using (a) the BL model and (b) the CONTRAfold model.

5. DISCUSSION

In this study, we have introduced a simple algorithm to transform a given initial base-pairing probability matrix (BPPM) into a new one that satisfies marginal probability constraints, by finding a BPPM that is “near” the initial one (where nearness is measured by the Kullback-Leibler distance). Although, strictly speaking, there is no guarantee that there exists a probability distribution on $\mathcal{S}(x)$ that produces the updated matrix, computational experiments showed that our algorithm is robust to missing data (cf. Fig. 3) and noise (cf. Fig. 4).

The proposed algorithms are applicable in any situation in which the base-pairing probability matrix of an RNA sequence can be found. This matrix can be computed by any algorithm, such as RNAfold (Hofacker et al., 1994), RNAstructure (Mathews et al., 1999) or CONTRAfold (Do et al., 2006). Notably, we also use our method with base-pairing probability matrices computed by Rfold (Kiryu et al., 2008) or RNAplfold (Bernhart et al., 2006) (these algorithms can restrict the maximum number of base pairs). This enables us to incorporate experimental data of very long sequences, such as the long transcriptome appearing in Underwood et al. (2010). It should also be emphasized that, by using an updated BPPM, many RNA informatics algorithms that internally employ BPPMs can be improved without modifying the algorithms themselves.

Converting experimental information given by SHAPE (Low and Weeks, 2010), PARS (Kertesz et al., 2010), and FragSeq (Underwood et al., 2010) technologies into marginal probability constraints (Definition 4) is not trivial. Some recent studies suggest that the experimental data include relatively large amount of noise, which can not be neglected (Kladwang et al., 2011). We believe that our algorithm including ε -relaxation is useful for those cases (cf. Fig. 4). In this article, we have mainly focused on the algorithms. Extensive computational experiments (using SHAPE, PARS, and FragSeq) will be undertaken in future work. It would be also interesting to investigate other ways of characterizing the distance (divergence) between BPPMs as alternatives to using the Kullback-Leibler distance (Amari and Nagaoka, 2000).

ACKNOWLEDGMENTS

MH is grateful to Ryoko Morioka and Koji Tsuda for valuable discussions. This work was inspired by their work related to macroeconomics (Morioka and Tsuda, 2011). This work was supported by MEXT KAKENHI (Grant-in-Aid for Young Scientists (A): 24680031).

DISCLOSURE STATEMENT

The author declares that no competing financial interests exist.

REFERENCES

- Amari, S., and Nagaoka, H. 2000. *Methods of Information Geometry. Translations of Mathematical Monographs* series, Vol. 191. Oxford University Press, New York.
- Andronescu, M., Condon, A., Hoos, H., et al. 2007. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*. 23, 19–28.
- Andronescu, M., Condon, A., Hoos, H.H., et al. 2010. Computational approaches for RNA energy parameter estimation. *RNA*. 16, 2304–2318.
- Bernhart, S.H., Hofacker, I.L., and Stadler, P.F. 2006. Local RNA base pairing probabilities in large sequences. *Bioinformatics*. 22, 614–615.
- Bregman, L. (1967). The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*. 7, 200–217.
- Carvalho, L. and Lawrence, C. (2008). Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3209–3214.
- Censor, Y.A. and Zenios, S.A. (1997). *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, New York.
- Deigan, K.E., Li, T.W., Mathews, D.H., et al. (2009). Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.* 106, 97–102.
- Do, C., Woods, D., and Batzoglou, S. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*. 22, e90–98.
- Esteller, M. (2011). Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12, 861–874.
- Gardner, P.P., Daub, J., Tate, J., et al. (2011). Rfam: Wikipedia, clans and the “decimal” release. *Nucleic Acids Res.* 39 (database issue), D141–145.
- Hamada, M., Tsuda, K., Kudo, T., et al. 2006. Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*. 22, 2480–2487.
- Hamada, M., Sato, K., Kiryu, H., et al. 2009a. CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics*. 25, 3236–3243.
- Hamada, M., Kiryu, H., Sato, K., et al. 2009b. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*. 25, 465–473.
- Hamada, M., Yamada, K., Sato, K., et al. 2011a. CentroidHomfold-LAST: accurate prediction of RNA secondary structure using automatically collected homologous sequences. *Nucleic Acids Res.* 39, W100–106.
- Hamada, M., Kiryu, H., Iwasaki, W., and Asai, K. 2011b. Generalized centroid estimators in bioinformatics. *PLoS ONE*. 6, e16450.
- Hamada, M., Sato, K., and Asai, K. 2011c. Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucleic Acids Res.* 39, 393–402.
- Hofacker, I., Fontana, W., Stadler, P., et al. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125, 167–188.
- Hofacker, I.L., Bernhart, S.H.F., and Stadler, P.F. 2004. Alignment of RNA base pairing probability matrices. *Bioinformatics*. 20, 2222–2227.
- Janssen, S., Schudoma, C., Steger, G., and Giegerich, R. 2011. Lost in folding space? Comparing four variants of the thermodynamic model for RNA secondary structure prediction. *BMC Bioinformatics*. 12, 429.
- Kato, Y., Sato, K., Hamada, M., et al. 2010. RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics*. 26, i460–466.
- Katoh, K., and Toh, H. (2008). Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics*. 9, 212.
- Kertesz, M., Wan, Y., Mazor, E., et al. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature*. 467, 103–107.
- Kiryu, H., Kin, T., and Asai, K. 2008. Rfold: an exact algorithm for computing local base pairing probabilities. *Bioinformatics*. 24, 367–373.
- Kladwang, W., VanLang, C.C., Cordero, P., and Das, R. 2011. Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry*. 50, 8049–8056.
- Kuhn, H., and Tucker, A. 1951. Nonlinear programming, 481–492. In J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, California.

- Leontis, N., and Westhof, E., eds. 2012. *RNA 3D Structure Analysis and Prediction (Nucleic Acids and Molecular Biology)*. Springer, New York.
- Low, J.T., and Weeks, K.M. 2010. SHAPE-directed RNA secondary structure prediction. *Methods*. 52, 150–158.
- Lu, Z.J., Gloor, J.W., and Mathews, D.H. 2009. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*. 15, 1805–1813.
- Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288, 911–940.
- McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*. 29, 1105–1119.
- Morioka, R., and Tsuda, K. 2011. Information geometry of input-output tables. *Technical note in IBISML*. 110, 161–167 (in Japanese).
- Novikova, I.V., Hennelly, S.P., and Sanbonmatsu, K.Y. 2012. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res.* 40, 5034–5051.
- Pang, P.S., Elazar, M., Pham, E.A., and Glenn, J.S. 2011. Simplified RNA secondary structure mapping by automation of SHAPE data analysis. *Nucleic Acids Res.* 39, e151.
- Quarrier, S., Martin, J.S., Davis-Neulander, L., et al. 2010. Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA*. 16, 1108–1117.
- Rivas, E., Lang, R., and Eddy, S.R. 2012. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA*. 18, 193–212.
- Sahraeian, S.M., and Yoon, B.J. 2011. PicXAA-R: efficient structural alignment of multiple RNA sequences using a greedy approach. *BMC Bioinformatics*. 12 Suppl 1, S38.
- Sato, K., Hamada, M., Mituyama, T., et al. 2009. A non-parametric bayesian approach for predicting RNA secondary structures. *Proceedings of the 9th International Conference on Algorithms in Bioinformatics*, 286–297.
- Sato, K., Kato, Y., Hamada, M., et al. 2011. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*. (27, i85–i93).
- Seemann, S., Gorodkin, J., and Backofen, R. 2008. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.* 36, 6355–6362.
- Seemann, S.E., Richter, A.S., Gesell, T., et al. 2011. PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*. 27, 211–219.
- Terai, G., Komori, T., Asai, K., and Kin, T. 2007. miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity. *RNA*. 13, 2081–2090.
- Underwood, J.G., Uzilov, A.V., Katzman, S., et al. 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*. 7, 995–1001.
- Wan, Y., Kertesz, M., Spitale, R.C., et al. 2011. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.* 12, 641–655.
- Washietl, S., Hofacker, I.L., Stadler, P.F., and Kellis, M. 2012. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res.* 40, 4261–4272.
- Watts, J.M., Dang, K.K., Gorelick, R.J., et al. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*. 460, 711–716.
- Wei, D., Alpert, L.V., and Lawrence, C.E. 2011. Rnag: A new gibbs sampler for predicting RNA secondary structure for unaligned sequences. *Bioinformatics*. 27, 2486–2493.
- Zakov, S., Goldberg, Y., Elhadad, M., and Ziv-Ukelson, M. 2011. Rich parameterization improves RNA structure prediction. *J. Comput. Biol.* 18, 1525–1542.

Address correspondence to:

Michiaki Hamada
Graduate School of Frontier Science
The University of Tokyo
5-1-5 Kashiwanoha
Kashiwa 277-8562
Japan

E-mail: mhamada@k.u-tokyo.ac.jp