BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 24 2012, pages 3218–3224
doi:10.1093/bioinformatics/bts612

Sequence analysis

Advance Access publication October 11, 2012

DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition

Kengo Sato[1,2,*], Yuki Kato[3], Tatsuya Akutsu[4], Kiyoshi Asai[2,5] and Yasubumi Sakakibara[1,2]

[1]Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223–8522, Japan, [2]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koto-ku, Tokyo 135–0064, Japan, [3]Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630–0192, Japan, [4]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611–0011, Japan and [5]Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, 5-1-5 Kashiwanoha, Chiba 277–8561, Japan

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: It is well known that the accuracy of RNA secondary structure prediction from a single sequence is limited, and thus a comparative approach that predicts a common secondary structure from aligned sequences is a better choice if homologous sequences with reliable alignments are available. However, correct secondary structure information is needed to produce reliable alignments of RNA sequences. To tackle this dilemma, we require a fast and accurate aligner that takes structural information into consideration to yield reliable structural alignments, which are suitable for common secondary structure prediction.

Results: We develop DAFS, a novel algorithm that simultaneously aligns and folds RNA sequences based on maximizing expected accuracy of a predicted common secondary structure and its alignment. DAFS decomposes the pairwise structural alignment problem into two independent secondary structure prediction problems and one pairwise (non-structural) alignment problem by the dual decomposition technique, and maintains the consistency of a pairwise structural alignment by imposing penalties on inconsistent base pairs and alignment columns that are iteratively updated. Furthermore, we extend DAFS to consider pseudoknots in RNA structural alignments by integrating IPknot for predicting a pseudoknotted structure. The experiments on publicly available datasets showed that DAFS can produce reliable structural alignments from unaligned sequences in terms of accuracy of common secondary structure prediction.

Availability: The program of DAFS and the datasets are available at http://www.ncrna.org/software/dafs/.

Contact: satoken@bio.keio.ac.jp

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

Received on May 23, 2012; revised on September 18, 2012; accepted on October 9, 2012

1 INTRODUCTION

Many functional RNAs form secondary structures that are related to their functions such as gene regulation and maturation of mRNAs, rRNAs and tRNAs. Therefore, analysis of RNA secondary structures can lead to unraveling their potential functions. Since experimental determination of RNA secondary structures is expensive and time consuming, computational prediction of RNA secondary structures has been frequently used. The most successful methods for predicting RNA secondary structures from a single sequence are based on minimizing free energy, such as Mfold (Zuker and Stiegler, 1981) and RNAfold (Hofacker, 2003). Alternative methods are based on probabilistic frameworks including stochastic context-free grammars that can model RNA secondary structures without pseudoknots (crossing base pairs on primary sequences). Several computational methods based on stochastic context-free grammars have been developed for modeling and analyzing non-coding RNAs (Do *et al.*, 2006; Eddy and Durbin, 1994; Rivas *et al.*, 2012; Sakakibara *et al.*, 1994; Sato *et al.*, 2010).

Nonetheless, the accuracy of RNA secondary structure prediction from a single sequence is known to be limited, and thus, a comparative approach that predicts a common secondary structure from aligned sequences is a better choice if homologous sequences with reliable alignments are available. Comparative methods based on alignment folding include RNAalifold (Bernhart *et al.*, 2008), Pfold (Knudsen and Hein, 2003), PETfold (Seemann *et al.*, 2008) and CentroidAlifold (Hamada *et al.*, 2011). However, we require correct secondary structure information to produce reliable alignments of RNA sequences in turn, which is a 'chicken-and-egg' problem.

To tackle this dilemma, Sankoff (1985) has proposed a dynamic programing (DP) algorithm for simultaneously aligning and folding RNA sequences whose computational complexity is $O(L^{3N})$ time and $O(L^{2N})$ space for N sequences of length L. Since a naïve implementation of the Sankoff algorithm is impractical because of unrealistic computational complexity, various algorithms have been developed to reduce the computational complexity of the Sankoff algorithm. Several pioneering studies including the first versions of FOLDALIGN (Gorodkin *et al.*, 1997) and Dynalign (Mathews and Turner, 2002) have implemented restricted Sankoff algorithms; FOLDALIGN disallows bifurcated structures, and Dynalign assumes the maximum distance of aligned nucleotides (note that later versions of both methods do not have such restrictions). Several succeeding methods have reduced the search space of the Sankoff algorithm by

*To whom correspondence should be addressed.

using precomputed posterior probabilities of base pairs and/or aligned columns (Do *et al*., 2008; Hofacker *et al*., 2004; Holmes, 2005; Kiryu *et al*., 2007; Will *et al*., 2007). Sampling-based approaches, instead of reducing the search space, have also been developed (Lindgreen *et al*., 2007; Meyer and Miklos, 2007; Xu *et al*., 2007). Ziv-Ukelson *et al*. (2010) have proposed a sparsification technique for the Sankoff algorithm that speeds up computation by a linear factor. Although various techniques for reducing the computation time have been developed, the Sankoff-style algorithms are still computationally expensive. Alternatively, several heuristic-based algorithms that do not strictly perform simultaneous aligning and folding of RNA sequences have been proposed such as sequence alignment with a structural scoring function (Dalli *et al*., 2006; Hamada *et al*., 2009a) and stem-based alignment (Perriquet *et al*., 2003; Tabei *et al*., 2008). LARA proposed by Bauer *et al*. (2007) formulates the problem of RNA structural alignment as a graph theoretical model for sequence alignment with additional simple structure information. Lagrangian relaxation is applied to relax the constraints for the structure information, then the problem is solved as sequence alignment with iterative updates of penalty scores. Note that LARA was not designed to predict common secondary structures since no folding model is explicitly considered.

In this article, we present a novel algorithm DAFS that performs Dual decomposition for Aligning and Folding RNA sequences Simultaneously. First, on the basis of the *maximizing expected accuracy* (MEA) principle, we design an objective function for a pairwise structural alignment as the expectation of the sum of the numbers of correctly predicted base pairs in a common secondary structure and those of correctly aligned columns in its pairwise alignment. Then, to maximize the objective function under several constraints for consistent pairwise structural alignment, we use the dual decomposition technique, which decomposes the pairwise structural alignment problem into two independent secondary structure prediction problems and one pairwise (non-structural) alignment problem. The algorithm maintains the consistency of a pairwise structural alignment by imposing penalties on inconsistent base pairs and alignment columns, and updates them iteratively by performing the Nussinov-style secondary structure prediction and the Needleman–Wunsch-style pairwise alignment with the penalized scoring function. We easily extend the algorithm to perform multiple alignment by applying a standard progressive alignment technique. Furthermore, we integrate the IPknot model (Sato *et al*., 2011) instead of the Nussinov model so that DAFS can simultaneously align and fold RNA sequences with pseudoknots. Major advantages of DAFS are summarized as follows: (i) DAFS is a fast and accurate implementation that simultaneously aligns and folds RNA sequences. Compared with the other Sankoff-style algorithms, DAFS achieves comparable accuracy, and even more importantly, runs much faster. Compared with the heuristic-based approaches such as sequence alignment with structural scores, DAFS is more accurate, especially in terms of common secondary structure prediction; (ii) DAFS is flexible and extensible because of the dual decomposition, which enables us to integrate state-of-the-art folding models such as the IPknot model for pseudoknotted secondary structures.

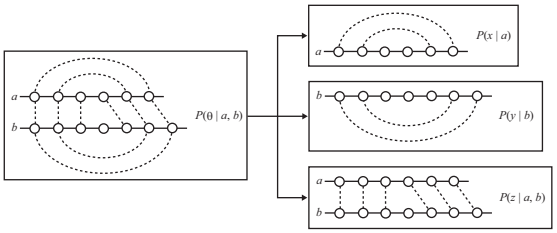

Fig. 1. An illustration of factorization of a probability distribution over a space of RNA structural alignments

2 METHODS

We present a novel method DAFS for simultaneous aligning and folding of RNA sequences by dual decomposition by which an integer programming (IP) problem for this task is solved. DAFS calculates a pairwise structural alignment, that is, a pairwise alignment that considers RNA secondary structures, by the following two steps when two unaligned sequences are given: (i) compute two base-pairing probability matrices and an alignment-matching probability matrix for the two sequences to factorize a probability distribution of structural alignments (Section 2.2, Fig. 1), and (ii) solve the IP problem (Section 2.3) of simultaneously aligning and folding the given sequences by dual decomposition (Section 2.4) to maximize the expected accuracy of the prediction. Then, the pairwise alignment algorithm is extended to compute a multiple alignment by a progressive alignment approach (Section 2.5). Furthermore, we describe an extension of our algorithm to perform RNA structural alignment that can consider pseudoknots (Section 2.6).

2.1 Preliminaries

Let $\Sigma = \{\mathrm{A,C,G,U}\}$ and Σ^* denote the set of all finite RNA sequences consisting of bases in Σ. For a sequence $a = a_1 a_2 \ldots a_n \in \Sigma^*$, let $|a|$ denote the number of symbols appearing in a, which is called the length of a. Given two RNA sequences $a, b \in \Sigma^*$, let $\mathcal{A}(a,b)$ be a set of all possible alignments of a and b, and $\mathcal{S}(a)$ be a set of all possible secondary structures of a. An alignment $z \in \mathcal{A}(a,b)$ is represented as a $|a| \times |b|$ binary-valued matrix $z = (z_{ik})$, where $z_{ik} = 1$ if and only if the base a_i is aligned with b_k. A secondary structure $x \in \mathcal{S}(a)$ is represented as a $|a| \times |a|$ binary-valued triangular matrix $x = (x_{ij})_{i<j}$, where $x_{ij} = 1$ if and only if bases a_i and a_j form a base pair. Let $\mathcal{SA}(a,b)$ be a set of all possible structural alignments of a and b. We write $\theta = (x,y,z)$, which means that a structural alignment $\theta \in \mathcal{SA}(a,b)$ consists of an alignment $z \in \mathcal{A}(a,b)$, and two secondary structures $x \in \mathcal{S}(a)$ and $y \in \mathcal{S}(b)$.

2.2 MEA-based scoring function

One of the most promising techniques to solve discrete high-dimensional problems on sequences such as RNA secondary structure prediction and RNA structural alignment is the MEA-based approach including centroid estimation (Carvalho and Lawrence, 2008; Hamada *et al*., 2009b).

Given unaligned RNA sequences, the goal of RNA structural alignment is to produce a reliable alignment, as well as to predict a reliable common secondary structure. To this end, we define a gain function of a structural alignment $\hat{\theta} = (\hat{x}, \hat{y}, \hat{z})$ with regard to the correct structural alignment $\theta = (x,y,z)$ as the weighted sum of gain functions $G_s(x,\hat{x})$ and $G_s(y,\hat{y})$ of respective secondary structures and a gain function $G_a(z,\hat{z})$ of the alignment as follows:

$$G(\theta,\hat{\theta}) = \alpha\{G_s(x,\hat{x}) + G_s(y,\hat{y})\} + G_a(z,\hat{z}) \quad (1)$$

where $\alpha > 0$ is a parameter that controls the weight of the secondary structures and the alignment. The gain function $G_s(x,\hat{x})$ of the secondary structure $\hat{x}$ is defined as follows:

$$G_s(x,\hat{x}) = (1-\tau)\,TP_s(x,\hat{x}) + \tau\,TN_s(x,\hat{x})$$

where $TP_s(x,\hat{x}) = \sum_{i<j} I(x_{ij}=1)I(\hat{x}_{ij}=1)$ is the number of true positive predictions, $TN_s(x,\hat{x}) = \sum_{i<j} I(x_{ij}=0)I(\hat{x}_{ij}=0)$ is the number of true negative predictions and $\tau \in [0,1]$ is a balancing parameter between true positives and true negatives. Here, $I(condition)$ is the indicator function that takes a value of 1 or 0 depending on whether the *condition* is true or false. The gain function $G_s(x,\hat{x})$ can be regarded as a kind of 'accuracy' that represents the weighted sum of the number of true predictions in $\hat{x}$, which is correlated to balanced accuracy measures such as Matthews correlation coefficient (MCC) and F-measure. Note that $G_s(x,\hat{x})$ is equivalent to the gain function of the γ-centroid estimator used in `CentroidFold` (Hamada *et al.*, 2009b) for $\tau = 1/(1+\gamma)$. Similarly, the gain function $G_a(z,\hat{z})$ of the alignment $\hat{z}$ is defined as the number of true predictions in $\hat{z}$ as follows:

$$G_a(z,\hat{z}) = (1-\sigma)\, TP_a(z,\hat{z}) + \sigma\, TN_a(z,\hat{z})$$

where $TP_a(z,\hat{z}) = \sum_{i,k} I(z_{ik}=1)I(\hat{z}_{ik}=1)$ is the number of true positive predictions and $TN_a(z,\hat{z}) = \sum_{i,k} I(z_{ik}=0)I(\hat{z}_{ik}=0)$ is the number of true negative predictions. Here, $\sigma \in [0,1]$ is a balancing parameter between true positives and true negatives.

In accordance with the MEA principle, we find a structural alignment $\hat{\theta}$ that maximizes the expectation of the gain function (1) under a given probability distribution over the space $\mathcal{SA}(a,b)$ of structural alignments:

$$\mathbb{E}_{\theta|a,b}[G(\theta,\hat{\theta})] = \sum\nolimits_{\theta\in\mathcal{SA}(a,b)} P(\theta|a,b)G(\theta,\hat{\theta}) \tag{2}$$

where $P(\theta|a,b)$ is a probability distribution of RNA structural alignments.

Unfortunately, computation of Equation (2) requires $O(|a|^3|b|^3)$ time and $O(|a|^2|b|^2)$ space even if we assume only pseudoknot-free secondary structures. Therefore, we assume independence of structure and alignment, and factorize the probability distribution of the structural alignments as follows:

$$P(\theta|a,b) \approx P(x|a)P(y|b)P(z|a,b) \tag{3}$$

where $P(x|a)$ and $P(y|b)$ are probability distributions of RNA secondary structures over $\mathcal{S}(a)$ and $\mathcal{S}(b)$, respectively, and $P(z|a,b)$ is a probability distribution of alignments over $\mathcal{A}(a,b)$ (Fig. 1). We can then approximate the expected gain function (2) by the following:

$$\begin{aligned}\mathbb{E}_{\theta|a,b}[G(\theta,\hat{\theta})] \approx{}& \sum\nolimits_{i,k}\left[p_{ik}^{(a,b)} - \sigma\right]\hat{z}_{ik} \\ &+ \alpha\left(\sum\nolimits_{i<j}\left[p_{ij}^{(a)} - \tau\right]\hat{x}_{ij} + \sum\nolimits_{k<l}\left[p_{kl}^{(b)} - \tau\right]\hat{y}_{kl}\right) + C\end{aligned} \tag{4}$$

where

$$p_{ij}^{(a)} = \sum_{x\in\mathcal{S}(a)} P(x|a)I(x_{ij}=1),$$
$$p_{kl}^{(b)} = \sum_{y\in\mathcal{S}(b)} P(y|b)I(y_{kl}=1)$$

are base-pairing posterior probability distributions,

$$p_{ik}^{(a,b)} = \sum_{z\in\mathcal{A}(a,b)} P(z|a,b)I(z_{ik}=1)$$

is an alignment-matching posterior probability distribution and C is a constant independent of $\hat{\theta} = (\hat{x},\hat{y},\hat{z})$ (see Section S1 in Supplementary Material for the derivation). If we assume pseudoknot-free structures for a and b, calculation of the posterior probabilities $p_{ij}^{(a)}$, $p_{kl}^{(b)}$ and $p_{ik}^{(a,b)}$ requires $O(|a|^3), O(|b|^3)$ and $O(|a||b|)$ in time and $O(|a|^2), O(|b|^2)$ and $O(|a||b|)$ in space, respectively.

2.3 IP formulation

Our objective is to find an RNA structural alignment that maximizes the approximate expected gain (4), satisfying the consistency of the RNA structural alignment. This optimization problem can be formulated as the following IP problem that maximizes:

$$\begin{aligned}S(x,y,z;a,b) = {}& \sum_{i,k}\left[p_{ik}^{(a,b)} - \sigma\right]z_{ik} \\ &+ \alpha\left(\sum_{i<j}\left[p_{ij}^{(a)} - \tau\right]x_{ij} + \sum_{k<l}\left[p_{kl}^{(b)} - \tau\right]y_{kl}\right)\end{aligned} \tag{5}$$

subject to:

$$\sum_{j<i} x_{ji} + \sum_{j>i} x_{ij} \le 1 \quad (1 \le \forall i \le |a|), \tag{6}$$

$$x_{ij} + x_{i'j'} \le 1 \quad (1 \le \forall i < \forall i' < \forall j < \forall j' \le |a|), \tag{7}$$

$$\sum_{l<k} y_{lk} + \sum_{l>k} y_{kl} \le 1 \quad (1 \le \forall k \le |b|), \tag{8}$$

$$y_{kl} + y_{k'l'} \le 1 \quad (1 \le \forall k < \forall k' < \forall l < \forall l' \le |b|), \tag{9}$$

$$\sum_{k} z_{ik} \le 1 \quad (1 \le \forall i \le |a|), \tag{10}$$

$$\sum_{i} z_{ik} \le 1 \quad (1 \le \forall k \le |b|), \tag{11}$$

$$z_{il} + z_{jk} \le 1 \quad (1 \le \forall i < \forall j \le |a|;\ 1 \le \forall k < \forall l \le |b|), \tag{12}$$

$$x_{ij} = \sum_{k<l} w_{ijkl} \quad (1 \le \forall i < \forall j \le |a|), \tag{13}$$

$$y_{kl} = \sum_{i<j} w_{ijkl} \quad (1 \le \forall k < \forall l \le |b|), \tag{14}$$

$$z_{ik} \ge \sum_{j<i,\, l<k} w_{jilk} + \sum_{j>i,\, l>k} w_{ijkl} \quad (1 \le \forall i \le |a|;\ 1 \le \forall k \le |b|) \tag{15}$$

where w_{ijkl} is a binary-valued variable such that $w_{ijkl} = 1$ if and only if a base pair (a_i, a_j) is aligned with a base pair (b_k, b_l). The constraint (6) means that each base a_i can be paired with at most one base (Fig. 2a), and the constraint (7) allows no pseudoknots in the secondary structure $x \in \mathcal{S}(a)$ (Fig. 2b). Similarly, the constraints (8) and (9) maintain the consistency of the secondary structure $y \in \mathcal{S}(b)$. The constraints (10) and (11) state that each base in a and b can be aligned with at most one base (Fig. 2c and d), and the constraint (12) allows no crossing match in the alignment $z \in \mathcal{A}(a,b)$ (Fig. 2e). The constraints (13)–(15) represent agreement between x, y and z that must be satisfied by the common secondary structure $\theta = (x,y,z) \in \mathcal{SA}(a,b)$. If we can drop the constraints (13)–(15), each component x, y and z of the structural alignment can be solved separately and efficiently. This means that the constraints (13)–(15) make the problem of RNA structural alignment extremely complex.

Consider the condition for $w_{ijkl} = 1$. Because of the constraints (13)–(15), $w_{ijkl} = 1$ implies $x_{ij} = y_{kl} = z_{ik} = z_{jl} = 1$. Therefore, to maximize the objective function (5), the contribution of $w_{ijkl} = 1$ should be positive:

$$[p_{ik}^{(a,b)} - \sigma] + [p_{jl}^{(a,b)} - \sigma] + \alpha\left([p_{ij}^{(a)} - \tau] + [p_{kl}^{(b)} - \tau]\right) > 0 \tag{16}$$

Furthermore, since the contribution of the base pair match, w_{ijkl} [the left-hand side of Equation (16)] should be larger than that of two loop matches z_{ik} and z_{jl}, that is, $[p_{ik}^{(a,b)} - \sigma] + [p_{jl}^{(a,b)} - \sigma]$, the following condition should be satisfied:

$$[p_{ij}^{(a)} - \tau] + [p_{kl}^{(b)} - \tau] > 0 \tag{17}$$

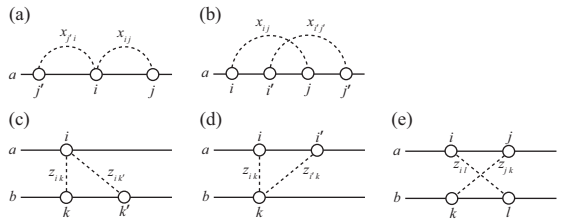

Fig. 2. An illustration of the constraints of the IP formulation. The diagrams (**a**) and (**b**) correspond to the constraints (6) and (7), respectively. Note that at most one variable shown by a broken curved line can take a value 1. The diagrams (**c**), (**d**) and (**e**) correspond to the constraints (10), (11) and (12), respectively

Therefore, it is no longer necessary to consider w_{ijkl} that does not satisfy the conditions (16) and (17). We call this the threshold cut technique, which accelerates our algorithm as we will describe in the next section.

2.4 Dual decomposition

Recall that the main difficulty of solving the IP problem of RNA structural alignment arises from the constraints (13)–(15). To circumvent this difficulty, we deal with these constraints using Lagrangian relaxation (Korte and Vygen, 2008). First, we define the Lagrangian dual by moving the constraints (13)–(15) to the objective function (5):

$$L(\lambda,\mu,\nu) = \max_{\substack{x\in\mathcal{S}(a),\, y\in\mathcal{S}(b),\\ z\in\mathcal{A}(a,b),\, w}} \Bigg\{ S(x,y,z;a,b) + \sum_{i<j}\lambda_{ij}\left(\sum_{k<l} w_{ijkl} - x_{ij}\right) + \sum_{k<l}\mu_{kl}\left(\sum_{i<j} w_{ijkl} - y_{kl}\right) + \sum_{i,k}\nu_{ik}\left(z_{ik} - \sum_{j<i,l<k} w_{jilk} - \sum_{j>i,l>k} w_{ijkl}\right)\Bigg\} \tag{18}$$

where $\lambda = \{\lambda_{ij} | i<j\}$, $\mu = \{\mu_{kl} | k<l\}$ and $\nu = \{\nu_{ik} | \nu_{ik} \geq 0\}$ are Lagrangian multipliers. We can then rewrite Equarion (18) as:

$$\begin{aligned} L(\lambda,\mu,\nu) = {} & \max_{x\in\mathcal{S}(a)} \sum_{i<j}\left[\alpha(p_{ij}^{(a)} - \tau) - \lambda_{ij}\right]x_{ij} \\ & + \max_{y\in\mathcal{S}(b)} \sum_{k<l}\left[\alpha(p_{kl}^{(b)} - \tau) - \mu_{kl}\right]y_{kl} \\ & + \max_{z\in\mathcal{A}(a,b)} \sum_{i,k}\left[p_{ik}^{(a,b)} - \sigma + \nu_{ik}\right]z_{ik} \\ & + \max_{w} \sum_{i<j,k<l}\left[\lambda_{ij} + \mu_{kl} - \nu_{ik} - \nu_{jl}\right]w_{ijkl} \end{aligned} \tag{19}$$

meaning that we can calculate each term of Equation (19) independently and efficiently by using the DP techniques: the Nussinov-style algorithm (Nussinov *et al.*, 1978) for the first and second terms and the Needleman–Wunsch-style algorithm (Needleman and Wunsch, 1970) for the third term; and by simply finding positive coefficients for the last term. This technique is called *dual decomposition* (Wainwright *et al.*, 2005).

Since the dual objective function $L(\lambda,\mu,\nu)$ gives an upper bound of the primal objective function (5), we aim to minimize Equation (19) with respect to the multipliers to obtain a better upper bound. The Lagrangian function $L(\lambda,\mu,\nu)$ is convex, but not differentiable (Korte and Vygen, 2008). Thus, to minimize the dual objective function (19), we can apply the subgradient optimization in which the Lagrangian multipliers λ_{ij}, μ_{kl} and ν_{ik} are iteratively updated by using their subgradients $\sum_{k<l} w_{ijkl} - x_{ij}$, $\sum_{i<j} w_{ijkl} - y_{kl}$ and $z_{ik} - \sum_{j<i,l<k} w_{jilk} - \sum_{j>i,l>k} w_{ijkl}$, respectively. As a result, we can obtain an algorithm similar to the gradient descent as shown in Figure 3, where $\eta^{(t)} > 0$ is a step size at each update. It

1: Calculate the posterior probabilities $p_{ij}^{(a)}$, $p_{kl}^{(b)}$ and $p_{ik}^{(a,b)}$.
2: Set $\lambda_{ij}^{(1)} = 0$, $\mu_{kl}^{(1)} = 0$ and $\nu_{ik}^{(1)} = 0$.
3: **for** $t = 1$ to T **do**
4: $\quad x^{(t)} \leftarrow \arg\max_{x\in\mathcal{S}(a)} \sum_{i<j}\left[\alpha(p_{ij}^{(a)} - \tau) - \lambda_{ij}^{(t)}\right]x_{ij}$
5: $\quad y^{(t)} \leftarrow \arg\max_{y\in\mathcal{S}(b)} \sum_{k<l}\left[\alpha(p_{kl}^{(b)} - \tau) - \mu_{kl}^{(t)}\right]y_{ij}$
6: $\quad z^{(t)} \leftarrow \arg\max_{z\in\mathcal{A}(a,b)} \sum_{i,k}\left[p_{ik}^{(a,b)} - \sigma + \nu_{ik}^{(t)}\right]z_{ik}$
7: $\quad w^{(t)} \leftarrow \arg\max_{w} \sum_{i<j,k<l}[\lambda_{ij}^{(t)} + \mu_{kl}^{(t)} - \nu_{ik}^{(t)} - \nu_{jl}^{(t)}]w_{ijkl}$
8: $\quad$ **if** $\theta^{(t)} = (x^{(t)}, y^{(t)}, z^{(t)})$ satisfies the constraints (13)–(15) **then**
9: $\quad\quad$ **return** $\theta^{(t)} = (x^{(t)}, y^{(t)}, z^{(t)})$
10: $\quad$ **end if**
11: $\quad \lambda_{ij}^{(t+1)} \leftarrow \lambda_{ij}^{(t)} - \eta^{(t)}\left(\sum_{k<l} w_{ijkl}^{(t)} - x_{ij}^{(t)}\right)$
12: $\quad \mu_{kl}^{(t+1)} \leftarrow \mu_{kl}^{(t)} - \eta^{(t)}\left(\sum_{i<j} w_{ijkl}^{(t)} - y_{kl}^{(t)}\right)$
13: $\quad \nu_{ik}^{(t+1)} \leftarrow \max\left\{0, \nu_{ik}^{(t)} - \eta^{(t)}\left(z_{ik}^{(t)} - \sum_{j<i,l<k} w_{jilk}^{(t)} - \sum_{j>i,l>k} w_{ijkl}^{(t)}\right)\right\}$
14: **end for**
15: **return** $\theta^{(T)} = (x^{(T)}, y^{(T)}, z^{(T)})$

Fig. 3. The algorithm for predicting RNA structural alignments using DD. T is the maximum number of iterations

has been proven that if $\lim_{t\to\infty}\eta^{(t)} = 0$ and $\sum_{t=1}^{\infty}\eta^{(t)} = \infty$, the Lagrangian dual $L(\lambda,\mu,\nu)$ always converges to the optimal value. The update is iterated until the solution is found or the number of iterations reaches the predefined maximum number of iterations T, which is sufficiently large.

The Lagrange multipliers can be regarded as penalty scores against inconsistency between alignments and secondary structures for the constraints (13)–(15).

The computational complexity of the algorithm in Figure 3 is dominated by that of the subproblems and the number of w_{ijkl}, which is at most $|a||b|(|a|+|b|)/2\tau$ by using the threshold cut technique described in Section 2.3. If we approximate $|a| \approx L$ and $|b| \approx L$, the time and space complexities of the algorithm are $O(TL^3/\tau)$ and $O(L^3/\tau)$, respectively. See Section S2 in Supplementary Material for details.

2.5 Extension to multiple alignment

We can easily extend pairwise alignment described in Section 2.4 to multiple alignment by the progressive approach (Feng and Doolittle, 1987) with the average base-pairing probability matrix and the average alignment-matching probability matrix.

Let A and B be two alignments of RNA sequences. We define the average base-pairing probabilities $p_{ij}^{(A)}$ to be the average of $p_{ij}^{(a)}$ over all sequences $a \in A$, remapped to the coordinates of the alignment A. Similarly, we define the average alignment-matching probabilities $p_{ik}^{(A,B)}$ to be the average of $p_{ik}^{(a,b)}$ over all pairs of sequences $a \in A$ and $b \in B$, remapped to the coordinates of the alignments A and B. We can predict the optimal structural alignment of two alignments A and B by the dual decomposition algorithm (Fig. 3) with $p_{ij}^{(A)}, p_{kl}^{(B)}$ and $p_{ik}^{(A,B)}$ instead of $p_{ij}^{(a)}, p_{kl}^{(b)}$ and $p_{ik}^{(a,b)}$.

Given a set of unaligned sequences, we first build a guide tree by clustering the sequences with a UPGMA using the expected accuracy similarity measure (Do *et al.*, 2005). We then perform progressive alignment by aligning alignments of the sequences according to the guide tree.

2.6 Extension to structural alignment with pseudoknots

As described in Section 2.4, our algorithm decomposes the master problem of RNA structural alignment into the three slave problems: two secondary structure prediction problems and one pairwise (non-structural) alignment problem, each of which can be solved independently. This allows us to easily take pseudoknots into account by replacing the Nussinov-style DP algorithm with the `IPknot` model (Sato *et al.*,

2011) for predicting secondary structures with pseudoknots. To the authors' knowledge, this is the first implementation that can simultaneously align and fold RNA sequences considering pseudoknotted secondary structures, except for SimulFold (Meyer and Miklos, 2007) and RNASampler (Xu *et al.*, 2007), which are based on a sampling algorithm. See Section S3 in Supplementary Material for details.

3 RESULTS

3.1 Implementation

Our algorithm was implemented as a program called DAFS. To calculate base-pairing probabilities, we used the McCaskill model (McCaskill, 1990) in the Vienna RNA package (Hofacker, 2003), using the free energy parameters estimated by the Boltzmann likelihood-based method (Andronescu *et al.*, 2010). To calculate alignment-matching probabilities, we used the ProbCons model (Do *et al.*, 2005). We implemented DAFS with two decoding algorithms for Lines 4 and 5 in Figure 3: the Nussinov decoding and the IPknot decoding. DAFS with the IPknot decoding can handle pseudoknotted secondary structures, whereas DAFS with the Nussinov decoding cannot.

Part of the IPknot decoding in DAFS was implemented using the CPLEX IP solver (http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/). In this study, we used empirically determined parameters: $\alpha = 4.0, \tau = 0.2, \sigma = 0.01$ and $T = 600$.

3.2 Datasets

For the benchmark, we created new datasets from Rfam 11.0 (Gardner *et al.*, 2011). Only manually curated seed alignments with consensus structures published in literature were used. We produced 691 pseudoknot-free alignments (PKfree dataset) and 82 pseudoknotted alignments (PK dataset). Each alignment contains 10 sequences from the Rfam families. Furthermore, to avoid selective bias in evaluation, we also used four datasets of RNA sequences that have been established in previous studies (Gardner *et al.*, 2005; Kiryu *et al.*, 2007; Lindgreen *et al.*, 2007; Tabei *et al.*, 2008). See Section S4 in Supplementary Material for more details.

3.3 Evaluation metrics

We evaluated the accuracy of predicted structural alignments through two measures for alignments and three measures for common secondary structures. To evaluate produced alignments, the following measures were used: (i) sum-of-pairs score (SPS) (Thompson *et al.*, 1999), which is defined as the proportion of correctly aligned matches in the predicted alignment, and (ii) structure conservation index (SCI) (Washietl *et al.*, 2005), which is defined as $\mathrm{SCI} = E_A/\bar{E}$, where E_A is the consensus minimum free energy computed by RNAalifold (Bernhart *et al.*, 2008) and $\bar{E}$ is the average minimum free energy over all RNA sequences in the predicted alignment. These measures can be calculated directly from produced alignments.

To evaluate predicted common secondary structures, we first mapped a predicted (reference, respectively) common secondary structure to each sequence according to the predicted (reference) alignment. Then, the predicted secondary structure mapped to each sequence was evaluated by the following three measures: (iii) sensitivity (SEN), which is defined as the proportion of correctly predicted base pairs in the reference secondary structure, i.e. $\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, (iv) positive predictive value (PPV), which is defined as the proportion of the correctly predicted base pairs in the predicted secondary structure, i.e. $\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, and (v) MCC, which is defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FN} \cdot \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

where TP is the number of base pairs appearing in both the reference and predicted structures, FP is the number of base pairs that appear in the predicted structure but not in the reference, TN is the number of base pairs that appear in neither structure and FN is the number of base pairs that appear in the reference but not in the predicted structure. Note that there exists a trade-off between SEN and PPV, and MCC is a balanced evaluation measure between them.

3.4 Effectiveness of dual decomposition

We first validated on the Murlet dataset (Kiryu *et al.*, 2007) that our algorithm can yield the optimal solution or a good approximation close to the optimal solution. Figure 4a shows the average number of constraints that are not satisfied at the maximum number of the iteration, T in Figure 3, ranging from 1 to 1000, indicating that satisfiable solutions can be achieved at $T \geq 600$. Figure 4b and c show elapsed time to process all the sequences in the dataset and accuracy, including SPS, SCI and MCC, at the maximum number of the iteration, respectively, comparing the results of the dual decomposition with the counterparts achieved by the CPLEX IP solver. These results clearly indicate that SPS, SCI and MCC sufficiently converged to the optimal at $T \geq 600$ even though the dual decomposition with $T = 600$ performed extremely faster than CPLEX. Therefore, we set $T = 600$ in subsequent experiments.

3.5 Structural alignments without pseudoknots

We conducted several computational experiments of predicting RNA structural alignments on the PKfree dataset, comparing DAFS with the following state-of-art aligners for structural alignment: (i) CentroidAlign version 1.00 (Hamada *et al.*, 2009a), (ii) RAF version 1.0 (Do *et al.*, 2008), (iii) LARA version 1.3.2a (Bauer *et al.*, 2007), and (iv) LocARNA version 1.6.1 (Will *et al.*, 2007). RAF and LocARNA, as well as DAFS, perform simultaneous aligning and folding of RNA sequences. CentroidAlign and LARA perform sequence alignment with structural scores. We also used the following standard alignment tool that does not consider any secondary structure as baselines in the comparison: (v) ProbConsRNA version 1.1 (Do *et al.*, 2005).

Table 1 and Supplementary Tables S2–S6 in Supplementary Material show the results on the datasets that do not contain pseudoknotted reference structures. These results indicate that DAFS achieved high accuracy in the measures for common secondary structures, especially in MCC that is a balanced measure between SEN and PPV, and is much faster than RAF, one of the most accurate competitors, although the accuracy of DAFS in the measures for alignments is not remarkable. Since accurate

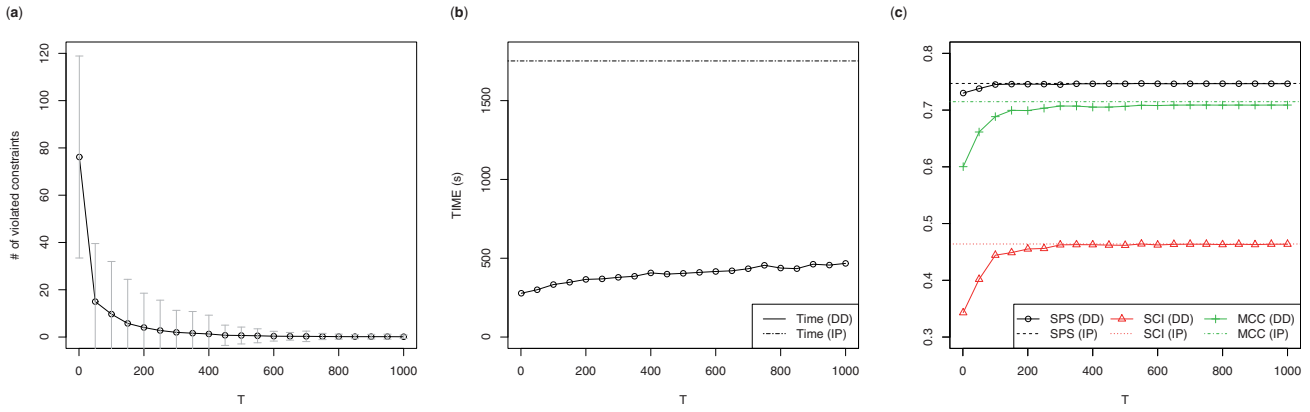

Fig. 4. The behavior of the dual decomposition algorithm as the maximum number of the iteration T is varied (DD = Dual Decomposition, IP = Integer Programming). (**a**) The mean and the standard deviation of the number of violated constraints over all the predictions at T. (**b**) The total execution time of DD and IP for all the alignments at T measured on a Linux workstation with Intel Xeon E5450 (3.0 GHz). (**c**) The accuracy of DD and IP at T

Table 1. The results on the PKfree dataset

Aligner	SPS	SCI	SEN	PPV	MCC	TIME
DAFS	0.82	0.61	**0.76**	0.77	**0.76**	2 275
CentroidAlign	**0.84**	0.61	0.68	**0.82**	0.73	1 010
RAF	0.82	0.58	0.73	0.79	0.75	11 693
LARA	0.81	0.62	0.69	0.80	0.73	29 734
LocARNA	0.79	**0.69**	0.70	0.79	0.73	63 935
ProbConsRNA	0.83	0.52	0.62	0.80	0.69	532

TIME means the total computation time in seconds measured on a Linux workstation with Intel Xeon E5450 (3.0 GHz). The bold values indicate the best score in each evaluation measure. For all the methods except DAFS, SEN, PPV and MCC, are calculated using common secondary structures predicted by CentroidAlifold with default parameters from produced alignments. See Section S4 in Supplementary Material for the results using RNAalifold and intrinsic predictions.

Table 2. The results on the PK dataset

Aligner	SPS	SCI	SEN	PPV	MCC	TIME
DAFS						
IPknot decoding	**0.80**	**0.57**	**0.61**	**0.70**	**0.64**	4 580
Nussinov decoding	**0.80**	0.56	0.54	0.67	0.60	674
RNASampler	0.70	0.55	0.52	0.69	0.59	164 206

TIME means the total computation time in seconds measured on a Linux workstation with Intel Xeon E5450 (3.0 GHz). The bold values indicate the best score in each evaluation measure. See Section S4 in Supplementary Material for the results using IPknot for DAFS with Nussinov decoding.

prediction of RNA secondary structures is crucial for unraveling potential functions of non-coding RNAs as mentioned in Section 1, fast and accurate structural alignments of DAFS shown in the results are preferable.

3.6 Structural alignments with pseudoknots

We conducted the experiment of predicting RNA structural alignments with pseudoknots on the PK dataset that contains pseudoknots in the reference structures, comparing DAFS with RNASampler version 1.3 (Xu *et al.*, 2007) with the option of '-X 1' that allows us to consider simple pseudoknots. Note that there is no existing practical structural aligner that can consider pseudoknots, except for RNASampler. Since RNASampler cannot recover common secondary structures with pseudoknots, we used IPknot with default parameters to predict common secondary structures from produced alignments. Table 2 clearly shows the advantage of DAFS with the IPknot decoding that can simultaneously align and fold RNA sequences with pseudoknots. Furthermore, the comparison between the Nussinov decoding and the IPknot decoding indicates that IPknot is successfully integrated into DAFS, meaning that DAFS is flexible and extensible because of the dual decomposition.

4 DISCUSSION

We developed a fast and accurate RNA structural aligner called DAFS that simultaneously aligns and folds RNA sequences by dual decomposition. Specifically, our method decomposes the RNA structural alignment problem into three subproblems, then respective problems are solved by simple and efficient algorithms such as the Nussinov-style DP and the Needleman–Wunsch-style DP. This means that DAFS can explicitly take the folding model into account, whereas LARA that has also used Lagrangian relaxation cannot. In fact, we showed that the dual decomposition technique enables us to integrate state-of-the-art folding models such as the IPknot model for pseudoknotted structures. This flexibility to adapt the folding models to our algorithm would lead us to further improvement by using an extended secondary structure model such as RNAwolf (zu Siederdissen *et al.*, 2011), where base-pairing probabilities for extended secondary structures are used.

One of the other differences between DAFS and LARA is the scoring system in the objective function. The scoring system of DAFS is based on the MEA principle, which has been successfully applied to various problems in bioinformatics (Do *et al.*, 2005,

2006; Hamada *et al*., 2009b; Hamada *et al*., 2011; Kato *et al*., 2010; Sato *et al*., 2011). The MEA-based scoring function not only improves accuracy in prediction but also makes DAFS run fast by the threshold cut technique, which is derived from the IP formulation based on the MEA-based scoring function described in Section 2.3. The threshold cut technique has also been applied to RactIP (Kato *et al*., 2010) and IPknot (Sato *et al*., 2011), and has contributed to their accuracy and efficiency.

Prediction accuracy of DAFS depends mainly on its scoring functions even though the method uses the approximate probability distribution for RNA structural alignments that assumes independence of alignment and structure. In fact, experimental results shown in Section S5 in Supplementary Material revealed that when we adopted the exact posterior probabilities for RNA structural alignments, a significant improvement was confirmed, though much computation time was spent on the predictions. Considering these results, there is room for further investigation into refinement of the scoring functions that make prediction accuracy compatible with practical running time.

ACKNOWLEDGEMENTS

The authors thank Dr Michiaki Hamada and our colleagues from the RNA Informatics Team at the Computational Biology Research Center (CBRC) for fruitful discussions. The super-computing resource was provided by Human Genome Center, Institute of Medical Science, University of Tokyo.

Funding: Grant-in-Aid for Young Scientists (B) (KAKENHI) from Japan Society for the Promotion of Science (22700305 to K.S. and 22700313 to Y.K.).

Conflict of Interest: none declared.

REFERENCES

Andronescu,M. *et al*. (2010) Computational approaches for RNA energy parameter estimation. *RNA*, **16**, 2304–2318.

Bauer,M. *et al*. (2007) Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinform*., **8**, 271.

Bernhart,S.H. *et al*. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinform*., **9**, 474.

Carvalho,L.E. and Lawrence,C.E. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl Acad. Sci. USA*, **105**, 3209–3214.

Dalli,D. *et al*. (2006) STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, **22**, 1593–1599.

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

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

Do,C.B. *et al*. (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, **24**, 68–76.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*., **22**, 2079–2088.

Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol*., **25**, 351–360.

Gardner,P.P. *et al*. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*., **33**, 2433–2439.

Gardner,P.P. *et al*. (2011) Rfam: Wikipedia, clans and the 'decimal' release. *Nucleic Acids Res*., **39**, D141–D145.

Gorodkin,J. *et al*. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*., **25**, 3724–3732.

Hamada,M. *et al*. (2009a) CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics*, **25**, 3236–3243.

Hamada,M. *et al*. (2009b) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465–473.

Hamada,M. *et al*. (2011) Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucleic Acids Res*., **39**, 393–402.

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

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

Holmes,I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinform*., **6**, 73.

Kato,Y. *et al*. (2010) RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics*, **26**, i460–466.

Kiryu,H. *et al*. (2007) Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, **23**, 1588–1598.

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

Korte,B. and Vygen,J. (2008) *Combinatorial Optimization: Theory and Algorithms*. Springer Verlag, Berlin, Germany.

Lindgreen,S. *et al*. (2007) MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics*, **23**, 3304–3311.

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

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

Meyer,I.M. and Miklos,I. (2007) SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol*., **3**, e149.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*., **48**, 443–453.

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

Perriquet,O. *et al*. (2003) Finding the common structure shared by two homologous RNAs. *Bioinformatics*, **19**, 108–116.

Rivas,E. *et al*. (2012) A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, **18**, 193–212.

Sakakibara,Y. *et al*. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res*., **22**, 5112–5120.

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

Sato,K. *et al*. (2010) A non-parametric Bayesian approach for predicting RNA secondary structures. *J. Bioinform. Comput. Biol*., **8**, 727–742.

Sato,K. *et al*. (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**, 85–93.

Seemann,S.E. *et al*. (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res*., **36**, 6355–6362.

Tabei,Y. *et al*. (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinform*., **9**, 33.

Thompson,J.D. *et al*. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*., **27**, 2682–2690.

Wainwright,M. *et al*. (2005) MAP estimation via agreement on trees: message-passing and linear programming. *IEEE Trans. Inf. Theory*, **51**, 3697–3717.

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

Will,S. *et al*. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol*., **3**, e65.

Xu,X. *et al*. (2007) RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics*, **23**, 1883–1891.

Ziv-Ukelson,M. *et al*. (2010) A faster algorithm for simultaneous alignment and folding of RNA. *J. Comput. Biol*., **17**, 1051–1065.

zu Siederdissen,C.H. *et al*. (2011) A folding algorithm for extended RNA secondary structures. *Bioinformatics*, **27**, i129–136.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*., **9**, 133–148.