

RNA Secondary Structure Prediction from Multi-Aligned Sequences

Michiaki Hamada

Abstract

It has been well accepted that the RNA secondary structures of most functional non-coding RNAs (ncRNAs) are closely related to their functions and are conserved during evolution. Hence, prediction of conserved secondary structures from evolutionarily related sequences is one important task in RNA bioinformatics; the methods are useful not only to further functional analyses of ncRNAs but also to improve the accuracy of secondary structure predictions and to find novel functional RNAs from the genome. In this review, I focus on common secondary structure prediction from a given *aligned* RNA sequence, in which one secondary structure whose length is equal to that of the input alignment is predicted. I systematically review and classify existing tools and algorithms for the problem, by utilizing the information employed in the tools and by adopting a unified viewpoint based on maximum expected gain (MEG) estimators. I believe that this classification will allow a deeper understanding of each tool and provide users with useful information for selecting tools for common secondary structure predictions.

Key words Common/consensus secondary structures, Comparative methods, Multiple sequence alignment, Covariation, Mutual information, Phylogenetic tree, Energy model, Probabilistic model, Maximum expected gain (MEG) estimators

1 Introduction

Functional non-coding RNAs (ncRNAs) play essential roles in various biological processes, such as transcription and translation regulation [11, 45, 49]. Not only nucleotide sequences but also secondary structures are closely related to the functions of ncRNAs, so secondary structures are conserved during evolution. Hence, the prediction of these conserved secondary structures (called “common secondary structure prediction” throughout this review) from evolutionarily related RNA sequences is among the most important tasks in RNA bioinformatics, because it provides useful information for further functional analysis of the targeted RNAs [4, 6, 28, 41, 46]. It should be emphasized that common secondary structure predictions are also useful to improve the accuracy of

secondary structure prediction [13, 48] or to find functional RNAs from the genome [15, 66].

Approaches to common secondary structure prediction are divided into two categories with respect to the input: (1) *unaligned* RNA sequences and (2) *aligned* RNA sequences. Common secondary structure predictions from *unaligned* RNA sequences can be solved by utilizing the Sankoff algorithm [54]. However, it is known that the Sankoff algorithm has huge computational costs: $O(L^{3n})$ and $O(L^{2n})$ for time and space, respectively, where L is the length of RNA sequences and n is the number of input sequences. For this reason, the second approach, whose input is *aligned* RNA sequences, is often used in actual analyses. The problem focused on this review is formulated as follows.

Problem 1 (RNA Common Secondary Structure Prediction of Aligned Sequences)

Given an input multiple sequence alignment A , predict a secondary structure y whose length is equal to the length of the input alignment. The secondary structure y is called the common (or consensus) secondary structure of the multiple alignment A .

In contrast to conventional RNA secondary structure prediction (see Fig. 1a), the input of common secondary structure prediction is a multiple sequence alignment of RNA sequences (see Fig. 1b) and the output is a secondary structure with the same length as the alignment. In general, the predicted common RNA secondary structure is expected to represent a secondary structure that commonly appears in the input alignments or is conserved during evolution.

To develop tools (or algorithms) for solving this problem, it should be made clear what a *better* common secondary structure is. However, the evaluation method for predicted common RNA secondary structures is *not* trivial. A predicted common secondary structure depends on not only RNA sequences in the alignment but also multiple alignments of input sequences, and in general, no reference (correct) common secondary structures are available (see **Note 1**). A predicted common secondary structure is therefore evaluated, based on reference RNA secondary structures for each *individual* RNA sequence in the alignment as follows.

Evaluation Procedure 1 (For Problem 1)

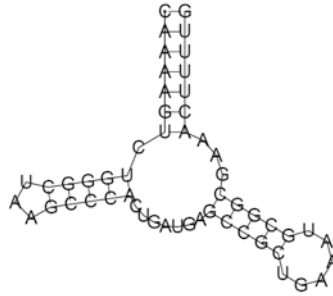
Given reference secondary structures for each RNA sequence, evaluate the predicted common secondary structure y as follows. First, map y onto each RNA sequence in the alignment A (see the column “Mapped structure with gaps” in Fig. 2, for example). Second, remove all gaps in each sequence and the corresponding base-pairs in the mapped secondary structure in order to maintain the consistency of the secondary structures (see the column “Mapped structure without gaps” in Fig. 2, for example). Third, calculate the quantities TP,

a Secondary structure prediction

CAAAAGUCUGGGCUAAGCCCACUGAUGAGCCGCUGAAAUGCGGCGAAACUUUUG

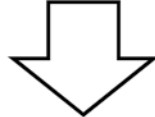


CAAAAGUCUGGGCUAAGCCCACUGAUGAGCCGCUGAAAUGCGGCGAAACUUUUG
(((((((.((((((...)))))) ((((((.....)))))) ...))))))



b Common secondary structure prediction

GGGCCUGUAGCUCAGAGGAUUAGAGCACGUGGCUACGAACCACGGUGUCGGGGUUCGAAUCCCUCCUCGCCCA
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCUGAUAAAGGGUGAGGUCGCGUGAUUCGAAUCCCUCCUCGCCCA
GGCGCCGUGGCGCAGUGGA--AGCGCGCAGGGCUCUAACCCUGAUGUCCUGGAUCGAAACCGAGCGGCGCUA
GCGUUGGUGGUUAGUGGUG-AGCAUAGCUGCCUCCAAAGCA-GUUGACCGGGUUCGAUUCGCGGCCAACGCA
ACUCCCUUAGUAUAUU---AAUAUAACUGACUUCUCCAAUUA-GUAGAUUCGAAU-AAACCCAGAAGAGAGUA



GGGCCCGUAGCACAGUGGA _AGAGCACAUGCCUCCAAACCA _GAUGUCCGGGUUCGAAUCCAGCCGAGCCCA
(((((((.((((((.....)))))) . ((((((.....)))))) ((((((.....))))))))) .

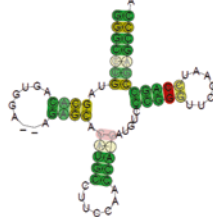


Fig. 1 (a) Conventional RNA secondary structure prediction, in which the input is an individual RNA sequence and the output is an RNA secondary structure of the sequence. **(b)** Common (or consensus) RNA secondary structure prediction in which the input is a multiple sequence alignment of RNA sequences and the output is an RNA secondary structure whose length is equal to the length of the alignment. The secondary structure is called the common (or consensus) secondary structure

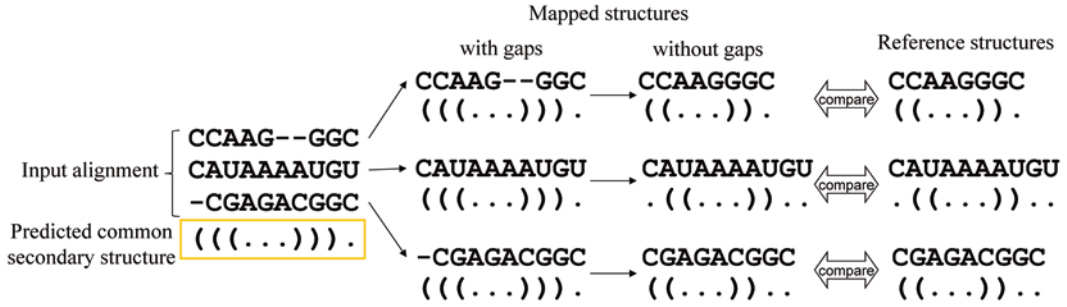


Fig. 2 An evaluation procedure for the predicted common RNA secondary structure of an input alignment, in which the reference secondary structure of each RNA sequence in the alignment is given. This procedure is based on the idea that a common secondary structure should reflect as many of the secondary structures of each RNA sequence in the input alignment as possible. Mapped RNA secondary structures without gaps are computed by getting rid of base-pairs that correspond to gaps. Note that it is difficult to compare a predicted common secondary structure with a *reference* common secondary structure, because the reference common RNA secondary structure for an arbitrary input alignment is not available in general. In most studies of common secondary structure prediction, evaluation is conducted by using this procedure or a variant. See [23] for a more detailed discussion of evaluation procedures for Problem 1

TN, FP, and FN for each mapped secondary structure $y^{(\text{map})}$ with respect to the reference secondary structure (TP, TN, FP, and FN are the respective numbers of true positive, true-negative, false-positive and false-negative base-pairs of a predicted secondary structure with respect to the reference structures). Finally, calculate the evaluation measures sensitivity (SEN), positive predictive value (PPV), and Matthew correlation coefficient (MCC) for the sum of TP, TN, FP, and FN over all the RNA sequences in the alignment A .

Figure 2 shows an illustrative example of Evaluation Procedure 1. Note that there exist a few variants of this procedure (e.g., [5]).

In this review, I aim to classify the existing tools (or algorithms) for Problem 1; These tools are summarized in Table 1, which includes all the tools for common secondary structure prediction as of 17th June 2013. To achieve this aim, I describe the information that is often utilized in common secondary structure predictions, and classify tools from a unified viewpoint based on maximum expected gain (MEG) estimators. I also explain the relations between the MEG estimators and Evaluation Procedure 1.

This review is organized as follows. In Subheading 2, I summarize the information that is commonly utilized when designing algorithms for common secondary structure prediction. In Subheading 3, several concepts to be utilized in the classification of tools are presented, and the currently available tools are classified within this framework.

Table 1
List of tools for common secondary structure prediction from *aligned* sequences

Tool	References	Description
(Without pseudoknot)		
CentroidAlifold	[19, 23]	Achieved superior performance in a recent benchmark (CompaRNA) [48]. Using an MEG estimator with the γ -centroid-type gain function (Subheading 3.3.2) in combination with a mixture probability distribution, including several types of information (Subheading 3.2.4)
ConStruct	[37, 72]	A semi-automatic, graphical tool based on mutual information (Subheading 2.2)
KNetFold	[6]	Computes a consensus RNA secondary structure from an RNA sequence alignment based on machine learning (Bayesian networks)
McCaskill-MEA	[32]	Adopting majority rule with the McCaskill energy model [40], leading to an algorithm that is robust to alignment errors
PETfold	[59]	Considers both phylogenetic information and thermodynamic stability by extending Pfold, in combination with an MEA estimation
Pfold	[34, 35]	Uses phylogenetic tree information with simple SCFG
PhyloRNAalifold	[14]	Incorporates the number of co-varying mutations on the phylogenetic tree of the aligned sequences into the covariance scoring of RNAalifold
PPfold	[63]	A multi-threaded implementation of the Pfold algorithm, which is extended to evolutionary analysis with a flexible probabilistic model for incorporating auxiliary data, such as data from structure probing experiments
RNAalifold	[5, 26, 27]	Considers both thermodynamic stability and co-variation in combination with RIBOSUM-like scoring matrices [33]
RSpredict	[62]	Takes into account sequence covariation and employs effective heuristics for accuracy improvement
(With pseudoknot)		
hxmatch	[73]	Computes consensus structures including pseudoknots based on alignments of a few sequences. The algorithm combines thermodynamic and covariation information to assign scores to all possible base-pairs, the base-pairs are chosen with the help of the maximum weighted matching algorithm
ILM	[52]	Uses mutual information and helix plot in combination with heuristic optimization
IPKnot	[57]	Uses MEG estimators with γ -centroid gains and heuristic probability distribution of RNA interactions together with integer linear programming to compute a decoded RNA secondary structure
MIfold	[12]	A MATLAB(R) toolbox that employs mutual information, or a related covariation measure, to display and predict common RNA secondary structure

To the best of my knowledge, this is a complete list of tools for the problem as of 17 June 2013. Note that tools for common secondary structure prediction from *unaligned* RNA sequences are not included in this list. See Table 2 for further details of the listed tools

Table 2
Comparison of tools (in Table 1) for common secondary structure predictions from aligned sequences

	Software ^a		Used information ^b							Gain ^c	Prob. dist. ^d
Name	SA	WS	TS	ML	CV	PI	MI	MR	EI		
(Without pseudoknot)											
CentroidAlifold	✓	✓	✓	✓	✓	✓		✓	^e	γ -cent	Any ^f
ConStruct	✓		✓		✓		✓	✓		na	na
KNetFold		✓	✓				✓			na	na
McCaskill-MEA	✓							✓		Contra	Av(Mc)
PETfold	✓	✓				✓				Contra	Pf+Av(Mc)
Pfold	✓			✓		✓				Delta	Pf
PhyloRNAalifold	✓				✓	✓				Delta	Ra
PPfold	✓			✓		✓			✓	Delta	Pf
RNAalifold	✓	✓	✓		✓					Delta	Ra
RSpredict	✓	✓	✓		✓					na	na
(With pseudoknot)											
hxmatch	✓		✓		✓					na	na
ILM	✓	✓	✓				✓			na	na
IPKnot	✓	✓	✓	✓	✓			✓		γ -cent	Any ^f
MIfold	✓ ^g						✓			na	na

^a Type of software available. *SA* stand alone, *WS* web server, *TS* thermodynamic stability (Subheading 2.1.1)
^b In the “Information used” columns, *ML* machine learning (Subheading 2.1.2); *CV* covariation (Subheading 2.3); *PI* phylogenetic (evolutionary) information (Subheading 2.4); *MI* mutual information (Subheading 2.2); *MR* majorityrule (Subheading 2.5); *EI* experimental information (Subheading 2.1.3)
^c In the column “Gain,” γ -cent: γ -centroid-type gain function (Subheading 3.3.2); contra: CONTRAfold-type gain function (Subheading 3.3.3)
^d In the column “Prob. dist.,” *Pf* Pfold model (Subheading 3.2.2); *Ra* RNAalipffold model (Subheading 3.2.1); *Av(Mc)* averaged probability distribution with McCaskill model (Subheading 3.2.3); “+” indicates a mixture distribution (of several models). “na” means “Not available” due to no use of probabilistic models
^e If the method proposed in [16] is used, experimental information derived from SHAPE [36] and PARS [31] is easily incorporated in CentroidAlign
^f CentroidAlifold and IPKnot can employ a mixed distribution given by an *arbitrary* combination of RNAalipffold, Pfold, and an averaged probability distribution based on the McCaskill or CONTRAfold models^gMATLAB codes are available

2 Materials

Several pieces of information are generally utilized in tools and algorithms for predicting common RNA ondary structures. These will now be briefly summarized.

2.1 Fitness to Each Sequence in the Input Alignment

The common RNA secondary structure should be a representative secondary structure among RNA sequences in the alignment. Therefore, the fitness of a predicted common secondary structure to *each* RNA sequence in the alignment is useful information. In particular, in Evaluation Procedure 1, the fitness of a predicted common secondary structure to each RNA sequence is evaluated.

This fitness is based on probabilistic models for RNA secondary structures of individual RNA sequence, such as the energy-based and machine learning models shown in Subheadings 2.1.1 and 2.1.2, respectively. These models provide a probability distribution of secondary structures of a given RNA sequence. ($p(\theta|x)$ denotes a probability distribution of RNA secondary structures for a given RNA sequence x .)

2.1.1 Thermodynamic Stability: Energy-Based Models

Turner's energy model [38] is an energy-based model, which considers the thermodynamic stability of RNA secondary structures. This model is widely utilized in RNA secondary structure predictions, in which experimentally determined energy parameters [38, 39, 75] are employed. In the model, structures with a lower free energy are more stable than those with a higher free energy. Note that Turner's energy model leads to a probabilistic model for RNA sequences, providing a probability distribution of secondary structures, which is called the McCaskill model [40].

2.1.2 Machine Learning (ML) Models

In addition to the energy-based models described in the previous subsection, probabilistic models for RNA secondary structures based on machine learning (ML) approaches have been proposed. In contrast to the energy-based models, machine learning models can automatically learn parameters from training data (i.e., a set of RNA sequences with secondary structures). There are several models based on machine learning which adopt different approaches: (1) Stochastic context free grammar (SCFG) models [10]; (2) the CONTRAfold model [9] (a conditional random field model); (3) the Boltzmann Likelihood (BL) model [1–3]; and (4) non-parametric Bayesian models [56].

See Rivas et al. [51] for detailed comparisons of probabilistic models for RNA secondary structures.

2.1.3 Experimental Information

Recently, experimental techniques to probe RNA structure by high-throughput sequencing (SHAPE [36]; PARS [31]; FragSeq [64]) have enabled genome-wide measurements of RNA structure. Those experimental techniques 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, such as HIV-1 [69], HCV (hepatitis C virus) [44], and large intergenic ncRNA (the steroid receptor RNA activator) [43], have been recently determined by combining those experimental techniques

with computational approaches. If available, such experimental information is useful in common secondary structure predictions, because they provide reliable secondary structures for each RNA sequence.

2.2 Mutual Information

The mutual information of the i th and j th columns in the input alignment is defined by

$$M_{ij} = \sum_{X,Y} f_{ij}(XY) \log \frac{f_{ij}(XY)}{f_i(X)f_j(Y)} = KL(f_{ij}(XY) || f_i(X)f_j(Y)) \quad (1)$$

where $f_i(X)$ is the frequency of base X at alignment position i ; $f_{ij}(XY)$ is the joint frequency of finding X in the i th column and Y in the j th column, and $KL(\cdot || \cdot)$ denotes the Kullback–Liebler distance between two probability distributions. As a result, the complete set of mutual information can be represented as an upper triangular matrix: $\{M_{ij}\}_{i < j}$.

Note that the mutual information score makes no use of base-pairing rules of RNA secondary structure. In particular, mutual information does not account for consistent non-compensatory mutations at all, although information about them would be useful when predicting common secondary structures as described in the next subsection.

2.3 Sequence Covariation of Base-Pairs

Because secondary structures of ncRNAs are related to their functions, mutations that preserve base-pairs (i.e., covariations of a base-pair) often occur during evolution. Figure 3 shows an example of covariation of base-pairs of tRNA sequences, in which many covariations of base-pairs are found, especially in the stem parts in the tRNA structure.

The covariation of the i th and j th columns in the input alignment is evaluated by the averaged number of compensatory mutations defined by

$$C_{ij} = \frac{2}{N(N-1)} \sum_{(x,y) \in A} d_{ij}^{x,y} \Pi_{ij}^x \Pi_{ij}^y \quad (2)$$

where N is the number of sequences in the input alignment A . For an RNA sequence x in A , $\Pi_{ij}^x = 1$ if x_i and x_j form a base-pair and $\Pi_{ij}^x = 0$ otherwise, and

$$d_{ij}^{x,y} = 2 - \delta(x_i, y_i) - \delta(x_j, y_j) \quad (3)$$

where δ is the delta function: $\delta(a, b) = 1$ only if $a = b$, and $\delta(a, b) = 0$ otherwise.

For instance, RNAalifold [65] uses the information of covariation in combination with the thermodynamic stability of common secondary structures.

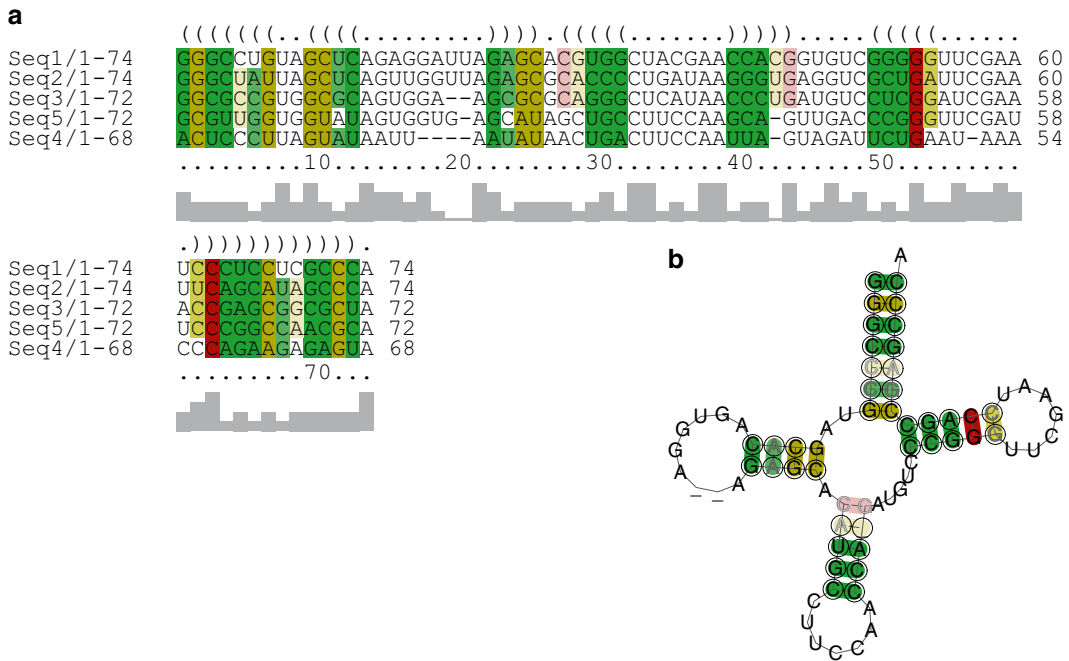


Fig. 3 An example of covariation of base-pairs in an alignment of tRNA sequences: **(a)** a multiple alignment of tRNA sequences and **(b)** a predicted common secondary structure of the alignment. The figures are taken from the output of an example on the RNAalifold [5] Web Server (<http://rna.tbi.univie.ac.at/cgi-bin/RNAalifold.cgi>)

2.4 Phylogenetic (Evolutionary) Information

Because most secondary structures of functional ncRNAs are conserved in evolution, the phylogenetic (evolutionary) information with respect to the input alignment is useful for predicting common secondary structures, and is employed by several tools. Pfold [34, 35] incorporates this information into probabilistic model for common secondary structures (Subheading 3.2.2) for the first time.

2.5 Majority Rule of Base-Pairs

Kiryu et al. [32] proposed the use of the majority rule of base-pairs in predictions of common secondary structures. This rule states that base-pairs supported by many RNA sequences should be included in a predicted common secondary structure. Specifically, Kiryu et al. utilized an averaged probability distribution of secondary structures among RNA sequences to predict a common RNA secondary structure from a given alignment (Subheading 3.2.3).

The aim of this approach is to mitigate alignment errors, because the effects of a minor alignment errors can be disregarded in the prediction of common secondary structures.

3 Methods

3.1 MEG Estimators

In this section, I classify the existing algorithms for common RNA secondary structure prediction (shown in Table 1) from a unified viewpoint, based on a previous study [23] in which the following type of estimator [18, 24] was employed.

$$\hat{y} = \operatorname{argmax}_{y \in S(A)} \sum_{\theta \in \mathcal{T}} G(\theta, y) p(\theta | A) \quad (4)$$

where $S(A)$ denotes a set of possible secondary structures with length $|A|$ (the length of the alignment A), $G(\theta, y)$ is called a “gain function” and returns a measure of the similarity between two common secondary structures, and $p(\theta | A)$ is a probabilistic distribution on $S(A)$. This type of estimator is an MEG estimator; When the gain function is designed according to accuracy measures for target problems, the MEG estimator is often called a “maximum expected accuracy (MEA) estimator” [18] (*see* Note 2).

In the following, a common secondary structure $\theta \in S(A)$ is represented as an upper triangular matrix $\theta = \{\theta_{ij}\}_{1 \leq i < j \leq |A|}$. In this matrix $\theta_{ij} = 1$ if the i th column in A forms a base-pair with the j th column in A , and $\theta_{ij} = 0$ otherwise.

The choices of $p(\theta | A)$ and $G(\theta, y)$ are described in Subheadings 3.2 and 3.3, respectively.

3.2 Choice of Probabilistic Models $p(\theta | A)$

The probabilities $p(\theta | A)$ provide a distribution of common RNA secondary structures given a multiple sequence alignment A . This distribution is given by the following models.

3.2.1 RNAalipffold Model

The RNAalipffold model is a probabilistic version of RNAalifold [27], which provides a probability distribution of common RNA secondary structures given a multiple sequence alignment. The distribution on $S(A)$ is defined by

$$p^{(\text{RNAalipffold})}(\theta | A) = \frac{1}{Z(T)} \exp\left(\frac{-E(\theta, A)}{kT} + \text{Cov}(\theta, A)\right) \quad (5)$$

where $E(\theta, A)$ is the averaged free energy of RNA sequences in the alignment with respect to the common secondary structure θ in A and $\text{Cov}(\theta, A)$ is the base covariation (cf. Subheading 2.3) with respect to the common secondary structure θ . The ML estimate of this distribution is equivalent to the prediction of RNAalifold.

Note that the negative part of the exponent in Eq. 5 is called the *pseudo* energy and it plays an essential role in finding ncRNAs from multiple alignments [15, 67].

3.2.2 Pfold Model

The Pfold model [34, 35] incorporates phylogenetic (evolutionary) information about the input alignments into a probabilistic distribution of common secondary structures:

$$p^{(\text{pfold})}(\theta | A) = p^{(\text{pfold})}(\theta | A, T, M) = \frac{p(A | \theta, T)p(\theta | M)}{p(A | T, M)} \quad (6)$$

where T is a phylogenetic tree, A is the input data (i.e., an alignment), M is a prior model for secondary structures (based on SCFGs; see **Note 3**). Unless the original phylogenetic tree T is obtained, T is taken to be the ML estimate of the tree, T^{ML} , given the model M and the alignment A .

3.2.3 Averaged Probability Distribution of Each RNA Sequence

Predictions based on RNAalifold and Pfold models tend to be affected by alignment errors in the input alignment. To address this, averaged probability distributions of RNA sequences involved in the input alignment were introduced by Kiryu et al. [32]. This leads to an MEG estimator with the probability distribution

$$p^{(\text{ave})}(\theta | A) = \frac{1}{n} \sum_{x \in A} p(\theta | x) \quad (7)$$

where $p(\theta | x)$ is a probabilistic model for RNA secondary structures, for example, the McCaskill model [40], the CONTRAfold model [9], the BL model [1–3], and others [56]. Note that neither covariation nor phylogenetic information about alignments is considered in this probability distribution.

In [32], the authors utilized the McCaskill model as a probabilistic model for individual RNA sequences (i.e., they used $p(\theta | x)$ in Eq. 14), and showed that their method was more robust with respect to alignment errors than RNAalifold and Pfold. Using averaged probability distributions for RNA sequences in an input alignment is compatible with Evaluation Procedure 1. See Hamada et al. [23] for a detailed discussion.

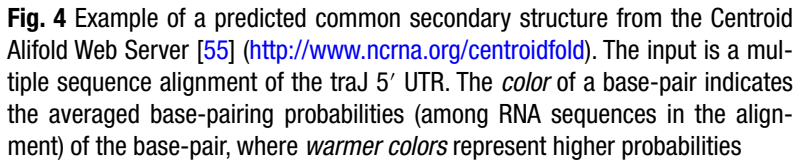
3.2.4 Mixture of Several Distributions

Hamada et al. [23] pointed out that arbitrary information can be incorporated into common secondary structure predictions by utilizing a *mixture* of probability distributions. For example, the probability distribution

$$p(\theta | A) = w_1 \cdot p^{(\text{pfold})}(\theta | A) + w_2 \cdot p^{(\text{alifold})}(\theta | A) + w_3 \cdot p^{(\text{ave})}(\theta | A), \quad (8)$$

where w_1 , w_2 , and w_3 are positive values that satisfy $w_1 + w_2 + w_3 = 1$, includes covariation (Subheading 2.3), phylogenetic tree (Subheading 2.4), and majority rule (Subheading 2.5) information.

In CentroidAlifold [23], users can employ a mixed distribution given by an arbitrary combination of RNAalifold (Subheading 3.2.1), Pfold (Subheading 3.2.2), and an averaged probability distribution



3.3 Choice of Gain Functions $G(\theta, y)$

A choice of the gain function in MEG estimators corresponds to the decoding method (i.e., prediction of one final common secondary structure from the distribution) of common RNA secondary structures, given a probabilistic model of common secondary structures.

3.3.1 The Kronecker Delta Function

A straightforward choice of the gain function is the Kronecker delta function:

$$G^{(\text{delta})}(\theta, y) (= \delta(\theta, y)) = \begin{cases} 1 & \text{if } y \text{ is the exactly same as } \theta \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The MEG estimator with this gain function is equivalent to the “maximum likelihood estimator” (ML estimator) with respect to a given probabilistic model for RNA common secondary structures (cf. Subheading 3.2), which predicts the secondary structure with the highest probability.

3.3.2 The γ -Centroid-Type Gain Function [19]

It is known that the probability of the ML estimation is extremely small, due to the immense number of secondary structures that could be predicted; this fact is known as the “uncertainty” of the solution, and often leads to issues in bioinformatics [17]. Because the MEG estimator with the delta function considers only the solution with the highest probability, it is affected by this uncertainty. A choice of gain function that partially overcomes this uncertainty of solutions is the γ -centroid-type gain function [19]:

$$G_{\gamma}^{(\text{centroid})}(\theta, y) = \sum_{i,j} \left[\gamma I(\theta_{ij} = 1) I(y_{ij} = 1) + I(\theta_{ij} = 0) I(y_{ij} = 0) \right] \quad (10)$$

where $\gamma > 0$ is a parameter that adjusts the relative importance of the SEN and PPV of base-pairs in a predicted structure (i.e., a larger γ produces more base-pairs in a predicted secondary structure). This gain function is motivated by the concept that more true base-pairs (TP and TN) and fewer false base-pairs (FP and FN) should be predicted [19] when the entire distribution of secondary structures is considered. Note that, when $\gamma = 1$, the gain function is equivalent to that of the centroid estimator [8].

3.3.3 The CONTRAfold-Type Gain Function [9]

In conventional RNA secondary structure predictions, another gain function has been proposed (see **Note 4**), which is based on the number of accurate predictions of every single position (*not* base-pair) in an RNA sequence (see **Note 5**). The CONTRAfold-type gain function is

$$G_{\gamma}^{(\text{contra})}(\theta, y) = \sum_{i=1}^{|A|} \left[\gamma \sum_{j:j \neq i} I(\theta_{ij}^* = 1) I(y_{ij}^* = 1) + \prod_{j:j \neq i} I(\theta_{ij}^* = 0) I(y_{ij}^* = 0) \right] \quad (11)$$

where θ^* and y^* are symmetric extensions of θ and y , respectively (i.e., $\theta_{ij}^* = \theta_{ij}$ for $i < j$ and $\theta_{ij}^* = \theta_{ji}$ for $j < i$). This gain function is also applicable to MEG estimators for common secondary structure predictions. It should be emphasized that MEG estimators based on the CONTRAfold-type gain function (for any $\gamma > 0$) do not include centroid estimators, while the γ -centroid-type gain function includes the centroid estimator as a special case (i.e., $\gamma = 1$).

3.3.4 Remarks About Choice of Gain Function

From a theoretical viewpoint, the γ -centroid-type gain function is more appropriate for Evaluation Procedure 1 than either the delta function or the CONTRAfold-type gain function (*see Note 6*), which is also supported by several empirical (computational) experiments. See Hamada et al. [23] for a detailed discussion.

Another choice of the gain function is MCC (or *F*-score), which takes a balance between SEN and PPV of base-pairs, and the MEG estimator with this gain function leads to an algorithm that maximizes *pseudo*-expected accuracy [22]. In addition, the estimator with this gain function includes only one parameter for predicting secondary structure (*see Note 7*).

3.4 Computation of Common Secondary Structure Through MEG Estimators

3.4.1 MEG Estimator with Delta Function

The MEG estimator with the delta function (that predicts the common secondary structure with the highest probability with respect to a given probabilistic model) can be computed by employing a CYK (Cocke–Younger–Kasami)-type algorithm. For example, *see* [65] for details.

3.4.2 MEG Estimator with γ -Centroid (or CONTRAfold) Type Gain Function

The MEG estimator with the γ -centroid-type (or CONTRAfold-type) gain function is computed based on “base-pairing probability matrices” (BPPMs) (*see Note 8*) and Nussinov-style dynamic programming (DP) [9, 23]:

$$M_{i,j} = \max \begin{cases} M_{i+1,j} \\ M_{i,j-1} \\ M_{i+1,j-1} + S_{ij} \\ \max_k [M_{i,k} + M_{k+1,j}] \end{cases} \quad (12)$$

where $M_{i,j}$ is the optimal score of the subsequence $x_{i\dots j}$ and S_{ij} is a score computed from the BPPM(s). For instance, for the γ -centroid estimator with the RNAalipffold model, the score S_{ij} is equal to $S_{ij} = (\gamma + 1)p_{ij}^{(\text{alipffold})} - 1$ where $p_{ij}^{(\text{alipffold})}$ is the base-pairing probability with respect to the RNAalipffold model. This DP algorithm maximizes the sum of (base-pairing) probabilities $p_{ij}^{(\text{alipffold})}$ which are larger than $1 / (\gamma + 1)$, and requires $O(|A|^3)$ time.

3.4.3 MEG Estimators with Averaged Probability Distributions

The MEG estimator with an averaged probability distribution (Subheading 3.2.3) can be computed by using averaged base-pairing probabilities, $\{p_{ij}\}_{i < j}$:

$$p_{ij}^{(\text{avc})} = \frac{1}{n} \sum_{x \in A} p_{ij}^{(x)} \quad (13)$$

where

$$p_{ij}^{(x)} = \begin{cases} \sum_{\theta \in S(x)} I(\theta_{\tau(i)\tau(j)} = 1) p(\theta | x') & \text{if both } x_i \text{ and } x_j \text{ are not gaps} \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

In the above, x' is the RNA sequence given by removing gaps from x and n is the number of sequences in the alignment A . The function $\tau(i)$ returns the position in x' corresponding to the position i in x .

The common secondary structure of MEG estimator with the γ -centroid gain function and the averaged probability distribution are computed by using the DP recursion in Eq. 12 where $S_{ij} = (\gamma + 1)p_{ij}^{(\text{ave})} - 1$. This procedure has a time complexity of $O(n|A|^3)$ where n is the number of sequences in the alignment.

3.4.4 MEG Estimators with a Mixture Distribution

The MEG estimator with a mixture of distribution (Subheading 3.2.4) and the delta function (Subheading 3.3.1) cannot be computed efficiently. However, if the γ -centroid-type (or CONTRAfold-type) gain function is utilized, the prediction can be conducted using a similar DP recursion to that in Eq. 12. For instance, the DP recursion of the γ -centroid-type gain function with respect to Eq. 8 is equivalent to the one in Eq. 12 with $S_{ij} = (\gamma + 1)p_{ij}^* - 1$ where

$$p_{ij}^* = w_1 \cdot p_{ij}^{(\text{pfold})} + w_2 \cdot p_{ij}^{(\text{alipffold})} + \frac{w_3}{n} \sum_{x \in A} p_{ij}^{(x)}. \quad (15)$$

In the above, $p_{ij}^{(\text{pfold})}$ and $p_{ij}^{(\text{alipffold})}$ are base-pairing probabilities for the Pfold and RNAalipffold models, respectively, and $\{p_{ij}^{(x)}\}$ is a base-pairing probability matrix with respect to a probabilistic model for secondary structures of single RNA sequence x (McCaskill or CONTRAfold model). Note that the total computational time of CentroidAlifold with a mixture of distributions still remains $O(n|A|^3)$.

3.4.5 MEG Estimators with Probability Distribution Including Pseudoknots

Using probability distributions of secondary structures with pseudoknots in MEG estimators generally has higher computational cost [42]. To overcome this, for example, IPKnot [57] utilizes an approximated method for determining the probability distribution as along with integer linear programming for predicting a final common secondary structure.

3.5 A Classification of Tools for Problem 1

Table 1 shows a comprehensive list of tools for common secondary structure prediction from aligned RNA sequences (in alphabetical order within groups that do or do not consider pseudoknots (see Note 9)). To the best of my knowledge, Table 1 is a complete list of tools for Problem 1 as of 17 June 2013.

In Table 2, the tools in Table 1 are classified based on the considerations of Subheadings 2 and 3. The classification leads to much useful information: (1) the pros and cons of each tool; (2) the similarity (or dissimilarity) among tools; (3) which tools are more suited to Evaluation Procedure 1; and (4) a unified framework within which to design algorithms for Problem 1. I believe that the classification will bring a deeper understanding of each tool, although several tools (which are not based on probabilistic models and depend fundamentally on heuristic approaches) cannot be classified in terms of MEG estimators.

3.6 Discussion

3.6.1 Multiple Sequence Alignment of RNA Sequences

Predicting a multiple sequence alignment (point estimation) from unaligned sequences is not reliable because the probability of the alignment becomes extremely small. This is called the “uncertainty” of alignments which raises serious issues in bioinformatics [17]. In one *Science* paper [74], for instance, the authors argued that the uncertainty of multiple sequence alignment greatly influences phylogenetic topology estimations: phylogenetic topologies estimated from multiple alignments predicted by five widely used aligners are different from one another. Similarly, point estimation of multiple sequence alignment will greatly affect consensus secondary structure prediction.

In Problem 1, because the quality of the multiple sequence alignment influences the prediction of common secondary structure, the input multiple alignment should be given by a multiple aligner which is designed specifically for RNA sequences. Although strict algorithms for multiple alignments taking into account secondary structures are equivalent to the Sankoff algorithm [54] and have huge computational costs, several multiple aligners which are fast enough to align long RNA sequences are available: these are CentroidAlign [20, 76], R-coffee [71], PicXXA-R [53], DAFS [58], and MAFFT [30]. In those multiple aligners, not only nucleotide sequences but also secondary structures are considered in the alignment, and they are, therefore, suitable for generating input multiple alignments for Problem 1.

Because the common secondary structure depends on multiple alignment, an approach adopted in RNAG [70] also seems promising. This approach iteratively samples from the conditional probability distributions $P(\text{Structure}|\text{Alignment})$ and $P(\text{Alignment}|\text{Structure})$. Note, however, that RNAG does not solve Problem 1 directly.

3.6.2 Improvement of RNA Secondary Structure Predictions Using Common Secondary Structure

Although several studies have been conducted for RNA secondary structure predictions for a single RNA sequence [9, 19, 38, 47], the accuracy is still limited, especially for long RNA sequences. By employing comparative approaches using homologous sequence information, the accuracy of RNA secondary structure prediction will be improved. In many cases, homologous RNA sequences of

the target RNA sequence are obtained, and someone would like to know the common secondary structure of those sequences. Gardner and Giegerich [13] introduced three approaches for comparative analysis of RNA sequences, and common secondary structure prediction is essentially utilized in the first of these. However, if the aim is to improve the accuracy of secondary structure predictions, common secondary structure prediction is not always the best solution, because it is not designed to predict the optimal secondary structure of a specific target RNA sequence. If you have a target RNA sequence for which the secondary structure is to be predicted, the approach adopted by the CentroidHomfold [21, 25] software is more appropriate than a method based on common RNA secondary structure prediction.

3.6.3 How to Incorporate Several Pieces of Information in Algorithms

As shown in this review, there are two ways to incorporate several pieces of information into an algorithm for common secondary structure prediction. The first approach is to modify the (internal) algorithm itself in order to handle the additional information. For example, PhyloRNAalifold [14] incorporates phylogenetic information into the RNAalifold algorithm by modifying the internal algorithm and PPfold [63] modifies the Pfold algorithm to handle experimental information. The drawbacks of this approach are the relatively large implementation cost and the heuristic combination of the information.

On the other hand, another approach adopted in CentroidAlifold [23] is promising because it can easily incorporate many pieces of information into predictions if a base-pairing probability matrix is available. Because the approach depends on only base-pairing probability matrices, and does not depend on the detailed design of the algorithm, it is easy to implement an algorithm using a mixture of distributions.

Moreover, a method to update a base-pairing probability matrix (computed using sequence information only) which incorporates experimental information [16] has recently been proposed. The method is independent of the probabilistic models of RNA secondary structures, and is suitable for incorporating experimental information into common RNA secondary structure prediction. A more sophisticated method by Washietl et al. [68] can also be used to incorporate experimental information into common secondary structure predictions, because it produces a BPPM that takes experimental information into account.

3.6.4 A Problem that Is Mathematically Related to Problem 1

Problem 1, which is considered in this paper, can be extended to predictions of RNA–RNA interactions, another important task in RNA bioinformatics (e.g., [29, 50]).

Problem 2 (Common Joint Structure Predictions of Two Aligned RNA Sequences)

Given two multiple alignments A_1 and A_2 of RNA sequences, then predict a joint secondary structure between A_1 and A_2 .

Because the mathematical structure of Problem 2 is similar to that of Problem 1, the ideas utilized in designing the algorithms for common secondary structure predictions can be adopted in the development of methods for this new problem. In fact, Seemann et al. [60, 61] have employed similar idea, adapting the PETfold algorithm to Problem 2 (implemented in the PETcofold software). Note that the problem of (pairwise) alignment between two multiple *alignments* (cf. see [24] for the details) has a similar mathematical structure to Problem 1.

3.7 Conclusion

In this review, I focused on RNA secondary structure predictions from *aligned* RNA sequences, in which a secondary structure whose length is equal to the length of the input alignment is predicted. A predicted common secondary structure is useful not only for further functional analyses of the ncRNAs being studied but also for improving RNA secondary structure predictions and for finding ncRNAs in genomes. In this review, I systematically classified existing algorithms on the basis of (1) the information utilized in the algorithms and (2) the corresponding MEG estimators, which consist of a gain function and a probability distribution of common secondary structures. This classification will provide a deeper understanding of each algorithm.

4 Notes

1. Reference common secondary structures are available for only reference multiple sequence alignments in the Rfam database [7] (<http://rfam.sanger.ac.uk/>).
2. The gain $G(\theta, y)$ is equal to the accuracy measure $\text{Acc}(\theta, y)$ for a prediction and references, so the MEG estimator maximizes the expected accuracy under a given probabilistic distribution.
3. The SCFG is based on the rules $S \rightarrow LS|L$, $F \rightarrow dFd|LS$, $L \rightarrow s|dFd$.
4. Historically, the CONTRAfold-type gain function was proposed earlier than the γ -centroid-type gain function.
5. This is not consistent with Evaluation Procedure 1, because accurate predictions of *base-pairs* with respect to reference structures are evaluated in it.
6. The CONTRAfold-type gain function has a bias toward accurate predictions of base-pairs, compared to the γ -centroid-type gain function.
7. The γ -centroid-type and CONTRAfold-type gain functions contain a parameter adjusting the ratio of SEN and PPV for a predicted secondary structure.

8. The BPPM is a probability matrix $\{p_{ij}\}$ in which p_{ij} is the marginal probability that the i th base x_i and the j th base x_j form a base-pair with respect to a given probabilistic distribution of secondary structures. For many probabilistic models, including the McCaskill model and the CONTRAfold model, the BPPM for a given sequence can be computed efficiently by utilizing inside–outside algorithms. *See* [40] for the details.
9. Tools for predicting common secondary structures *without* pseudoknots are much faster than those for predicting secondary structures *with* pseudoknots.

Acknowledgement

This work was supported in part by MEXT KAKENHI (Grant-in-Aid for Young Scientists (A): 24680031; Grant-in-Aid for Scientific Research (A): 25240044).

References

1. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics* 23(13):19–28
2. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP (2010) Computational approaches for RNA energy parameter estimation. *RNA* 16(12):2304–2318
3. Andronescu MS, Pop C, Condon AE (2010) Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA* 16(1):26–42
4. Balik A, Penn AC, Nemoda Z, Greger IH (2013) Activity-regulated RNA editing in select neuronal subfields in hippocampus. *Nucl Acids Res* 41(2):1124–1134
5. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinform* 9:474
6. Bindewald E, Shapiro BA (2006) RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA* 12(3):342–352
7. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A (2013) Rfam 11.0: 10 years of RNA families. *Nucl Acids Res* 41(Database issue):D226–D232
8. Carvalho LE, Lawrence CE (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc Natl Acad Sci USA* 105(9):3209–3214
9. Do CB, Woods DA, Batzoglou S (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* 22(14):e90–e98
10. Dowell RD, Eddy SR (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinform* 5:71
11. Esteller M (2011) Non-coding RNAs in human disease. *Nat Rev Genet* 12(12):861–874
12. Freyhult E, Moulton V, Gardner P (2005) Predicting RNA structure using mutual information. *Appl Bioinform* 4(1):53–59
13. Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinform* 5:140
14. Ge P, Zhang S (2013) Incorporating phylogenetic-based covarying mutations into RNAalifold for RNA consensus structure prediction. *BMC Bioinform* 14(1):142
15. Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF (2010) Rnaz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput* 15:69–79
16. Hamada M (2012) Direct updating of an RNA base-pairing probability matrix with marginal probability constraints. *J Comput Biol* 19(12):1265–1276
17. Hamada M (2014) Fighting against uncertainty: an essential issue in bioinformatics. *Briefings Bioinform* 15(5):748–767
18. Hamada M, Asai K (2012) A classification of bioinformatics algorithms from the viewpoint

- of maximizing expected accuracy (MEA). *J Comput Biol* 19(5):532–549
19. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics* 25(4):465–473
 20. Hamada M, Sato K, Kiryu H, Mituyama T, Asai K (2009) CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics* 25(24):3236–3243
 21. Hamada M, Sato K, Kiryu H, Mituyama T, Asai K (2009) Predictions of RNA secondary structure by combining homologous sequence information. *Bioinformatics* 25(12):i330–i338
 22. Hamada M, Sato K, Asai K (2010) Prediction of RNA secondary structure by maximizing pseudo-expected accuracy. *BMC Bioinform* 11:586
 23. Hamada M, Sato K, Asai K (2011) Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucl Acids Res* 39(2):393–402
 24. Hamada M, Kiryu H, Iwasaki W, Asai K (2011) Generalized centroid estimators in bioinformatics. *PLoS ONE* 6(2):e16450
 25. Hamada M, Yamada K, Sato K, Frith MC, Asai K (2011) CentroidHomfold-LAST: accurate prediction of RNA secondary structure using automatically collected homologous sequences. *Nucl Acids Res* 39(Web Server issue):W100–W106
 26. Hofacker IL (2007) RNA consensus structure prediction with RNAalifold. *Methods Mol Biol* 395:527–544
 27. Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. *J Mol Biol* 319(5):1059–1066
 28. Jager D, Pernitzsch SR, Richter AS, Backofen R, Sharma CM, Schmitz RA (2012) An archaeal sRNA targeting cis- and trans-encoded mRNAs via two distinct domains. *Nucl Acids Res* 40(21):10964–10979
 29. Kato Y, Sato K, Hamada M, Watanabe Y, Asai K, Akutsu T (2010) RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics* 26(18):i460–i466
 30. Katoh K, Toh H (2008) Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinform* 9:212
 31. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467(7311):103–107
 32. Kiryu H, Kin T, Asai K (2007) Robust prediction of consensus secondary structures using averaged base-pairing probability matrices. *Bioinformatics* 23(4):434–441
 33. Klein RJ, Eddy SR (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinform* 4:44
 34. Knudsen B, Hein J (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* 15(6):446–454
 35. Knudsen B, Hein J (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucl Acids Res* 31(13):3423–3428
 36. Low JT, Weeks KM (2010) SHAPE-directed RNA secondary structure prediction. *Methods* 52(2):150–158
 37. Luck R, Graf S, Steger G (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucl Acids Res* 27(21):4208–4217
 38. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288(5):911–940
 39. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA* 101(19):7287–7292
 40. McCaskill JS (1990) The equilibrium partition function and base-pair binding probabilities for RNA secondary structure. *Biopolymers* 29(6–7):1105–1119
 41. Meer EJ, Wang DO, Kim S, Barr I, Guo F, Martin KC (2012) Identification of a cis-acting element that localizes mrna to synapses. *Proc Natl Acad Sci* 109(12):4639–4644
 42. Nebel ME, Weinberg F (2012) Algebraic and combinatorial properties of common RNA pseudoknot classes with applications. *J Comput Biol* 19(10):1134–1150
 43. Novikova IV, Hennelly SP, Sanbonmatsu KY (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucl Acids Res* 40(11):5034–5051
 44. Pang PS, Elazar M, Pham EA, Glenn JS (2011) Simplified RNA secondary structure mapping by automation of SHAPE data analysis. *Nucl Acids Res* 39(22):e151
 45. Pauli A, Rinn JL, Schier AF (2011) Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* 12(2):136–149

46. Penn AC, Balik A, Greger IH (2013) Steric antisense inhibition of AMPA receptor Q/R editing reveals tight coupling to intronic editing sites and splicing. *Nucl Acids Res* 41(2):1113–1123
47. Proctor JR, Meyer IM (2013) COFOLD: an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucl Acids Res* 41(9):e102
48. Puton T, Kozłowski LP, Rother KM, Bujnicki JM (2013) CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucl Acids Res* 41(7):4307–4323
49. Qureshi IA, Mehler MF (2012) Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nat Rev Neurosci* 13(8):528–541
50. Richter AS, Backofen R (2012) Accessibility and conservation: general features of bacterial small RNA–mRNA interactions? *RNA Biol* 9(7):954–965
51. Rivas E, Lang R, Eddy SR (2012) A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA* 18(2):193–212
52. Ruan J, Stormo GD, Zhang W (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* 20(1):58–66
53. Sahraeian SM, Yoon BJ (2011) PicXAA-R: efficient structural alignment of multiple RNA sequences using a greedy approach. *BMC Bioinform* 12(Suppl 1):S38
54. Sankoff D (1985) Simultaneous solution of the RNA folding alignment and protosequence problems. *SIAM J Appl Math* 45:810–825
55. Sato K, Hamada M, Asai K, Mituyama T (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucl Acids Res* 37(Web Server issue):W277–W280
56. Sato K, Hamada M, Mituyama T, Asai K, Sakakibara Y (2010) A non-parametric bayesian approach for predicting rna secondary structures. *J Bioinform Comput Biol* 8(4):727–742
57. Sato K, Kato Y, Hamada M, Akutsu T, Asai K (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics* 27(13):85–93
58. Sato K, Kato Y, Akutsu T, Asai K, Sakakibara Y (2012) DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition. *Bioinformatics* 28(24):3218–3224
59. Seemann SE, Gorodkin J, Backofen R (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucl Acids Res* 36(20):6355–6362
60. Seemann SE, Menzel P, Backofen R, Gorodkin J (2011) The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences. *Nucl Acids Res* 39(Web Server issue):W107–W111
61. Seemann SE, Richter AS, Gesell T, Backofen R, Gorodkin J (2011) PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics* 27(2):211–219
62. Spirollari J, Wang JT, Zhang K, Bellofatto V, Park Y, Shapiro BA (2009) Predicting consensus structures for RNA alignments via pseudo-energy minimization. *Bioinform Biol Insights* 3:51–69
63. Sukosd Z, Knudsen B, Kjems J, Pedersen CN (2012) PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics* 28(20):2691–2692
64. Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, Lowe TM, Salama SR, Haussler D (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods* 7(12):995–1001
65. Washietl S, Hofacker IL (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol* 342(1):19–30
66. Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* 23(11):1383–1390
67. Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA* 102(7):2454–2459
68. Washietl S, Hofacker IL, Stadler PF, Kellis M (2012) RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucl Acids Res* 40(10):4261–4272
69. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Swanstrom R, Burch CL, Weeks KM (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 460(7256):711–716
70. Wei D, Alpert LV, Lawrence CE (2011) RNAG: a new Gibbs sampler for predicting RNA secondary structure for unaligned sequences. *Bioinformatics* 27(18):2486–2493
71. Wilm A, Higgins DG, Notredame C (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucl Acids Res* 36(9):e52

72. Wilm A, Linnenbrink K, Steger G (2008) ConStruct: improved construction of RNA consensus structures. *BMC Bioinform* 9:219
73. Witwer C, Hofacker IL, Stadler PF (2004) Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans Comput Biol Bioinform* 1(2):66–77
74. Wong KM, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. *Science* 319(5862):473–476
75. Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base-pairs. *Biochemistry* 37(42):14719–14735
76. Yonemoto H, Asai K, Hamada M (2013) CentroidAlign-Web: a fast and accurate multiple aligner for long non-coding RNAs. *Int J Mol Sci* 14(3):6144–6156

RNA Bioinformatics

Picardi, E. (Ed.)

2015, XII, 415 p. 102 illus., 73 illus. in color., Hardcover

ISBN: 978-1-4939-2290-1

A product of Humana Press