BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 15 2007, pages 1883–1891
doi:10.1093/bioinformatics/btm272

Sequence Analysis

RNA Sampler: A new sampling based algorithm for common RNA secondary structure prediction and structural alignment

Xing Xu, Yongmei Ji[†] and Gary D. Stormo[*]

Department of Genetics, Washington University, School of Medicine, St. Louis, MO 63110, USA.

Received on February 9, 2007; revised on May 10, 2007; accepted on May 13, 2007

Advance Access publication May 30, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Non-coding RNA genes and RNA structural regulatory motifs play important roles in gene regulation and other cellular functions. They are often characterized by specific secondary structures that are critical to their functions and are often conserved in phylogenetically or functionally related sequences. Predicting common RNA secondary structures in multiple unaligned sequences remains a challenge in bioinformatics research.

Methods and Results: We present a new sampling based algorithm to predict common RNA secondary structures in multiple unaligned sequences. Our algorithm finds the common structure between two sequences by probabilistically sampling aligned stems based on stem conservation calculated from intrasequence base pairing probabilities and intersequence base alignment probabilities. It iteratively updates these probabilities based on sampled structures and subsequently recalculates stem conservation using the updated probabilities. The iterative process terminates upon convergence of the sampled structures. We extend the algorithm to multiple sequences by a consistency-based method, which iteratively incorporates and reinforces consistent structure information from pairwise comparisons into consensus structures. The algorithm has no limitation on predicting pseudoknots. In extensive testing on real sequence data, our algorithm outperformed other leading RNA structure prediction methods in both sensitivity and specificity with a reasonably fast speed. It also generated better structural alignments than other programs in sequences of a wide range of identities, which more accurately represent the RNA secondary structure conservations.

Availability: The algorithm is implemented in a C program, RNA Sampler, which is available at http://ural.wustl.edu/software.html.

Contact: xingxu@ural.wustl.edu and stormo@genetics.wustl.edu

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

1 INTRODUCTION

Non-coding RNAs (ncRNA) and RNA regulatory motifs play important roles in gene regulation and other cellular functions (Eddy, 2001; Stormo and Ji, 2001; Bartel, 2004; Winkler, 2005). They are often characterized by evolutionarily conserved secondary structures that are critical to their functions. Identification of correct RNA secondary structures will help elucidate the functions of the RNAs and finding conserved RNA secondary structures in related sequences will provide promising candidates for novel ncRNA genes or RNA regulatory motifs.

RNA secondary structure prediction is one of the most challenging problems in bioinformatics and has been a topic of intense study. Predicting the RNA secondary structure of a single sequence relies on the idea that the functional RNA structure is the thermodynamically most stable or near-optimal structure. Mfold (Zuker, 1994) and RNAfold (Hofacker, *et al.*, 1994) employ dynamic programming to compute the RNA secondary structure with minimum free energy (MFE) on a single sequence. By considering energy density, length normalized free energy, Densityfold (Alkan, *et al.*, 2006) delocalizes the thermodynamic cost of forming an RNA substructure and improves on secondary structure prediction via free energy minimization. PKNOTS (Rivas and Eddy, 1999) extends the Zuker algorithm to optimal RNA pseudoknot structure prediction. However, optimal structures that are predicted based on current energy parameters do not necessarily represent the real structures.

A more reliable approach is to use comparative analysis to predict consensus RNA secondary structures from multiple related sequences. One strategy is to align related sequences first and then fold the alignment. Mutual information (Gutell, *et al.*, 1992), probabilistic covariance models (Eddy and Durbin, 1994) and stochastic context-free grammars (SCFG) (Sakakibara, *et al.*, 1994; Knudsen and Hein, 1999; 2003) have been effectively used to detect and model complementary covariations that are indicative of conserved base pairing interactions to predict common structures. RNAalifold (Hofacker, *et al.*, 2002) incorporates both thermodynamic stability and sequence covariation to predict common structures from alignments. Maximum weighted matching (MWM), a graph-theoretical approach, was introduced to predict common secondary structures allowing pseudoknots (Cary and Stormo, 1995; Tabaska, *et al.*, 1998). Reliable structural alignments are usually prerequisites for these algorithms, in which sequences can neither be too similar to provide covariance information nor too divergent to provide a reliable alignment.

Another strategy proposed by Sankoff (1985) is to simultaneously align and fold RNA sequences. However, the computational complexity of the Sankoff algorithm is too high to be practical even for two sequences. By making simplifications and applying restrictions, Foldalign (Gorodkin, *et al.*, 2001) and Dynalign

[*]To whom correspondence should be addressed.

[†]Present address: Rosetta Inpharmatics LLC, a wholly owned subsidiary of Merck & Co., Inc., Seattle, WA 98109, USA.

(Mathews and Turner, 2002) make the Sankoff algorithm practical on short sequences but are still slow. Stemloc (Holmes and Rubin, 2002; Holmes, 2005), PMcomp (Hofacker, *et al.*, 2004), RNAscf (Bafna, *et al.*, 2006) and CARNAC (Touzet and Perriquet, 2004) combine ideas, such as SCFG, RNA base pairing matrix alignment and stem comparison, with the Sankoff algorithm to simultaneously achieve common structures and structural alignments. These approaches usually predict common structures of two sequences first, and then use a progressive method like that of ClustalW (Thompson, *et al.*, 1994) to find common structures shared by multiple sequences.

The third strategy is to predict the structures of individual sequences separately and then align the structures. It has been implemented in programs such as RNAshapes (Giegerich, *et al.*, 2004; Steffen, *et al.*, 2006) and MARNA (Siebert and Backofen, 2005). comRNA (Ji, *et al.*, 2004) uses a different scheme that applies graph-theoretical approaches to compare and find stems conserved across multiple sequences first and then assembles conserved stem blocks to form consensus structures, in which pseudoknots are allowed.

We present a new algorithm for predicting common structures in multiple unaligned sequences. It adopts the stem assembly idea from comRNA (Ji, *et al.*, 2004) and combines intrasequence base pairing probabilities and intersequence base alignment probabilities to measure stem conservation. It employs an iterative procedure to probabilistically sample compatible aligned stem pairs and update the probabilities based on sampled structures until convergence. We extend the algorithm to multiple sequences by a consistency-based method (Do, *et al.*, 2005). Our algorithm has no limitation on predicting pseudoknots. In extensive testing on real sequence data, it outperformed other leading programs in both sensitivities and specificities with a reasonably fast speed.

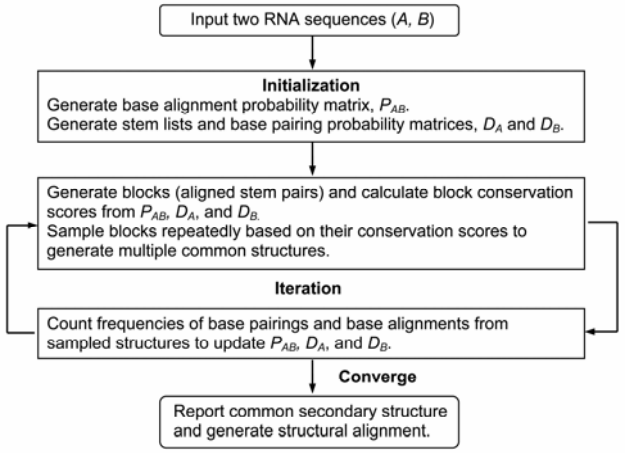

Fig. 1. Flow-chart of the sampling algorithm on two RNA sequences.

2 METHODS

RNA secondary structures are formed by base pairing stacks (stems) and single stranded regions (loops). Predicting common RNA secondary structures on multiple sequences can be simplified to finding compatible conserved stems between sequences. A pair of conserved stems are assumed to not only have high chances to form base pairings in individual sequences but also align well between sequences. Therefore, we combine the intrasequence base pairing probabilities and the intersequence base alignment probabilities to measure stem conservation.

Our algorithm repeatedly generates multiple common structures by probabilistically sampling compatible stem pairs based on their conservation. It updates the base pairing and base alignment probabilities by counting frequencies of actual base pairs and base alignments in the sampled structures, and uses the updated probabilities to recalculate stem conservation and resample structures. This procedure is iterated until the sampled structures converge. We extend our algorithm to multiple sequences by using a consistency-based method to iteratively incorporate and reinforce consistent structure information from pairwise sequence comparisons into structures common to multiple sequences.

The overall scheme of the core algorithm in finding common structures in two sequences is illustrated in Fig. 1. It solves the problem in two major steps: initialization and iteration.

2.1 Initialization step

2.1.1 Calculate base alignment probabilities $P_{AB}(a_i, b_k)$ is defined as the base alignment probability (or match probability) between base a_i from sequence A and base b_k from sequence B. We calculate the initial base alignment probabilities of all possible pairs of bases between two sequences by using the partition function based method (Miyazawa, 1995; Muckstein, *et al.*, 2002) (see Supplementary Appendix I).

2.1.2 Calculate base pairing probabilities $D_A(a_i, a_j)$ (or $D_B(b_k, b_l)$) is defined as the base pairing probability between bases a_i and a_j (or b_k and b_l) in sequence A (or B). By default, pseudoknots are not allowed. The initial base pairing probabilities for all possible complementary base pairs in each sequence can be either calculated by RNAfold (Hofacker, *et al.*, 1994), an implementation of the partition function based algorithm (McCaskill, 1990), or approximated by the posterior probabilities generated by CONTRAfold (Do, *et al.*, 2006). These two initialization methods have almost the same performance in our tests, and both are available as options for users. In this paper, we use RNAfold in all tests. We also designed a fast heuristic sampling method to calculate the initial base pairing probabilities that allow pseudoknot structures (see Supplementary Appendix II).

2.1.3 Generate a list of all possible stems for each sequence A stem is defined as consecutive A·U, G·C or G·U base pairs with a minimum length of *sl*, where *sl* is a user definable parameter, usually 3 (default) or 4 (for long sequences or sequences with pseudoknot structures). No internal loop or bulge is allowed in a stem. We list all possible stems within each sequence. These stems may overlap or be independent of each other and various combinations can form different RNA secondary structures.

2.2 Iteration step

2.2.1 A block is a pair of aligned stems We can generate a series of alignments between two stems by sliding one stem along the other stem, in which consecutive base pairs from one stem are aligned to those from the other. An alignment consists of two corresponding halves with equal widths: the 5' arm and the 3' arm, as shown in Fig. 2.

Given an alignment (Φ) between two stems (u and v) (see Fig. 2), we calculate the conservation score as:

$$W(\Phi \mid u, v) = \sum_{\substack{(a_i, a_j)\in u;\\ (b_k, b_l)\in v;\\ (a_i, b_k) \text{ and } (a_j, b_l)\in \Phi}} \left[P_{AB}(a_i, b_k) \cdot D_A(a_i, a_j) \cdot P_{AB}(a_j, b_l) \cdot D_B(b_k, b_l) \right]$$

where (a_i, a_j) and (b_k, b_l) are complementary base pairs in stem u from sequence A and stem v from sequence B, respectively; (a_i, b_k) and (a_j, b_l) are aligned bases in the 5' arm and 3' arm, respectively; $P_{AB}(a_i, b_k)$ and $P_{AB}(a_j, b_l)$ are the intersequence base alignment probabilities; $D_A(a_i, a_j)$ and $D_B(b_k, b_l)$ are the intrasequence base pairing probabilities.

A *block* (β) is defined as the alignment of two stems (u and v) that gives the highest conservation score (W):

$$\beta = \arg\max_{\Phi}[W(\Phi \mid u, v)]$$

The conservation score collectively measures both sequence and structure conservation of the stem pairs, which are represented by the base alignment probabilities and base pairing probabilities, respectively. Only those blocks consisting of conserved stem pairs would have high conservation scores.

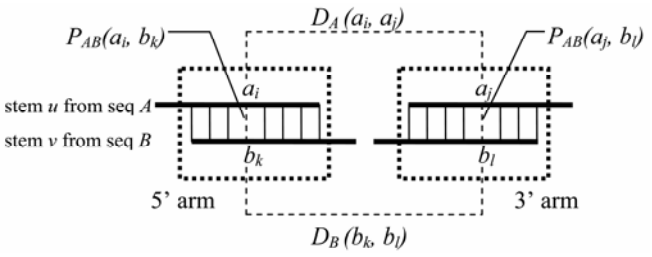

Fig. 2. Structure of a block. The alignment of two stems consists of consecutive aligned base pairs. (a_i, a_j) and (b_k, b_l) are complementary base pairs in stem u from sequence A and stem v from sequence B, respectively. (a_i, b_k) and (a_j, b_l) are aligned bases in the 5' arm and 3' arm, respectively. $P_{AB}(a_i, b_k)$ and $P_{AB}(a_j, b_l)$ are the intersequence base alignment probabilities; $D_A(a_i, a_j)$ and $D_B(b_k, b_l)$ are the intrasequence base pairing probabilities. The 5' and 3' aligned arms of two stems in dotted boxes that give the highest conservation score, W, constitute a *block*.

2.2.2 Generate a list of blocks A complete list of blocks is generated by aligning every stem in one sequence to every stem in the other sequence. To reduce search space, only those blocks consisting of stem pairs with reciprocal best conservation scores are considered. For instance, the block consisting of stem u in sequence A and stem v in sequence B has the highest conservation score among all combinations between stem u and any stem from B, and also among all combinations between stem v and any stem from A. To further reduce computational complexity, a user definable parameter d is introduced, which is the maximum shift distance allowed between aligned positions in two stems.

2.2.3 Sample compatible blocks to generate common structures A probabilistic sampling approach is used to pick compatible blocks based on their conversation scores and assemble them into common structures. The chance for a block (β) to be sampled is defined by the probability:

$$p(\beta) = \frac{W(\beta)}{\sum_{\beta \in \{all\ blocks\}} W(\beta)}$$

The higher the conservation score of a block, the more likely it is picked to be part of a structure (see Supplementary Appendix III for details).

This sampling process is repeated S times, where S is the sample size, and S common structures are ultimately generated in each iteration.

2.2.4 Iteratively update the base pairing probabilities and base alignment probabilities We calculate the frequencies of base pairs appearing in those S common structures for each sequence to approximate the new base pairing probabilities:

$$D_A^r(a_i, a_j) = \frac{D_A^{r-1}(a_i, a_j) \cdot T + \sum_s C_s^{AB}(a_i, a_j)}{T + S}$$

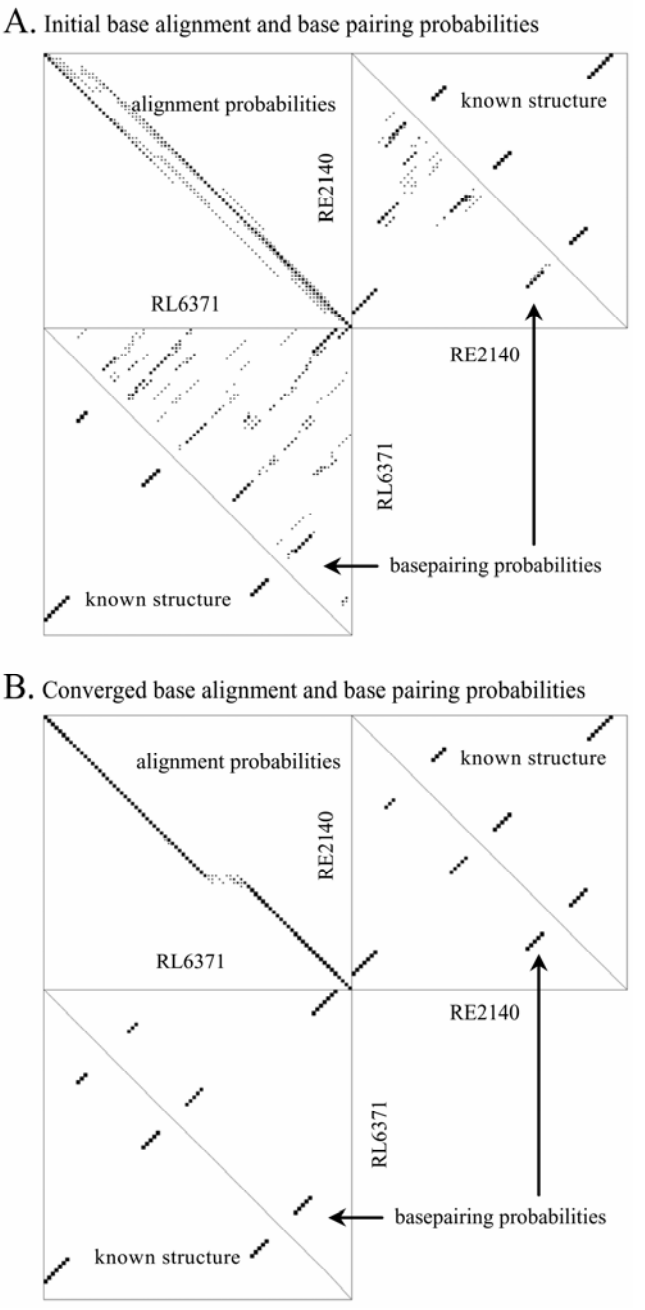

Fig. 3. Illustration of the initial and converged base alignment and base pairing probability matrices between two tRNA sequences, RL6371 and RE2140. In each panel, the upper left square represents the base alignment probability matrix between the two sequences. The other two squares represent the base pairing probability matrices of each sequence respectively. The area of a dot in the matrices corresponds to the value of probability, ranging from 0 to 1. The base pairing probability matrix is symmetrical to its diagonal, so only half of the matrix is drawn, and the known structure of the sequence is drawn on the other half. A). Initial probability matrices. B). Converged probability matrices.

where $s \in \{S$ sampled structures between sequence A and $B\}$; given any s,

$$C_s^{AB}(a_i,a_j)=\begin{cases}1 & (a_i,a_j)\ forms\ a\ base\ pair\ in\ s;\\ 0 & (a_i,a_j)\ does\ not\ form\ a\ base\ pair\ in\ s.\end{cases}$$

r is the iteration number; $D_A^{r-1}(a_i, a_j)\cdot T$ is the pseudocount portion corresponding to the base pairing probabilities from the last iteration. T is a user definable parameter for the total number of pseudocounts, which can be 1, $Sqrt(S)$, S or $10\%\cdot S$ (default) depending on how much the user wants the probabilities from the last iteration to affect the subsequent iteration.

We calculate the frequencies of base alignments in these S common structures between two sequences to approximate the new base alignment probabilities:

$$P_{AB}^{r}(a_i,b_k)=\frac{P_{AB}^{r-1}(a_i,b_k)\cdot T+\sum_s C_s^{AB}(a_i,b_k)}{T+S}$$

where $s\in\{S$ sampled structures between sequence A and $B\}$; given any s,

$$C_s^{AB}(a_i,b_k)=\begin{cases}1 & a_i\ and\ b_k\ are\ aligned\ in\ a\ stem\ region\ of\ s;\\ p(a_i,b_k) & a_i\ and\ b_k\ are\ in\ the\ same\ aligned\ loop\ region\ of\ s;\\ 0 & all\ other\ cases.\end{cases}$$

$p(a_i, b_k)$ is the recalculated alignment probability of (a_i, b_k) only if a_i and b_k are in the same loop region of the sampled structure; $P_{AB}^{r-1}(a_i, b_k)\cdot T$ is the pseudocount portion corresponding to the alignment probabilities from the last iteration.

Steps 2.2.2, 2.2.3 and 2.2.4 are repeated for multiple iterations. The updated base pairing and base alignment probabilities are used to recalculate the conservation scores of blocks and their probabilities to be sampled in the subsequent iteration. Through iterations, the blocks that are formed by conserved stems obtain intensified conservation scores and have increasingly higher chances to be sampled, while other blocks tend to have their conservation scores attenuated. The sampled structures tend to converge to sets of compatible conserved blocks.

We examined the convergence of sampled structures on 55 two-sequence sets generated from pairwise combinations of 11 tRNA sequences used in previous studies (Sprinzl, *et al.*, 1998). In each iteration, 100 structures were sampled. The number of unique sampled structures decreased quickly after a few iterations for all sequence sets, dropping by 2~16-fold after 2 iterations and converging after 4 ~ 7 iterations (see Supplementary Fig. 1). Fig. 3 shows the initial and converged base alignment and base pairing probabilities between two tRNA sequences, RL6371 and RE2140. The initial base alignment and base pairing probability matrices shown in Fig. 3A can result in ambiguous structures and alignments between the two sequences. However, as shown in Fig. 3B, just after a few iterations, the base pairing and base alignment probabilities of most stems were attenuated or diminished, while only a few stems that matched known structures were intensified, indicating the convergence of sampled structures.

2.3 Report the common structure in two sequences

A greedy algorithm is used to assemble converged blocks into the final consensus structure: the block with the highest conservation score is picked and any conflicting block is deleted or trimmed; this process is repeated until no blocks remain. The chosen blocks are compatible with each other and form the backbone of the common structure. We then use ClustalW to realign single-stranded regions separately, and assemble them with conserved blocks to generate the final structural alignment.

2.4 From two sequences to multiple sequences

We extend the algorithm from taking two RNA sequences as input to multiple (N) sequences. In the initialization step, the base alignment probabilities between each pair of sequences and the base pairing probabilities for each sequence are calculated using the same procedure as described above. In the iteration step, we sample common structures between all pairwise sequences and apply a consistency-based method to update probabilities and recalculate conservation scores, which incorporate information from all sampled structures into the subsequent iteration of structure sampling on pairwise sequences.

In each iteration, S structures are sampled between each pair of sequences, and each sequence is involved in $(N\text{-}1)\cdot S$ sampled structures.

The base pairing probabilities of a sequence (A) incorporate information from all pairwise common structures between A and any other sequence (B), which are calculated as:

$$D_A^r(a_i,a_j)=\frac{1}{(N-1)}\cdot\sum_B\left(\frac{D_A^{r-1}(a_i,a_j)\cdot T+\sum_s C_s^{AB}(a_i,a_j)}{T+S}\right)$$

where $s\in\{S$ sampled structures between A and $B\}$; $B\in\{N$ sequences$\}$, $B\neq A$; $C_s^{AB}(a_i, a_j)$ and T are the same as those defined previously for two sequences.

The base alignment probabilities between any pair of sequences are calculated using the same procedure as described above for two sequences.

At the end of each iteration, the conservation scores of all blocks and their probabilities to be sampled in the next iteration are updated using the newly calculated base alignment and base pairing probabilities. The iteration process continues until the number of unique structures sampled between every pair of sequences converges.

2.5 Report the common structure in multiple sequences

Upon convergence of sampled structures after iterations, the final common structure shared by multiple sequences is reported by assembling sets of compatible conserved stems using a greedy algorithm. All blocks between all pairwise sequences are sorted based on their final conservation scores, and the block with the highest score becomes the seed; any conflicting block is deleted or trimmed, and all blocks that contain one of the stems in the seed are collected; the new stem from the block with the highest conservation score in this subset is added to the seed. This process is repeated until no blocks remain in this subset. The stems in the seed now correspond to one common stem shared by multiple sequences. The next block with the highest conservation score among the remaining blocks becomes the next seed to obtain another set of common stems. This procedure is repeated until no blocks remain. We then use ClustalW to align conserved stems and single-stranded regions separately, and assemble them together to generate the final structural alignment.

3 RESULTS

Our algorithm was implemented in a C program called RNA Sampler. We tested it on various data sets containing two or multiple real sequences, and compared its performance to that of the other existing RNA secondary structure prediction programs representing different methodologies, including CARNAC (Touzet and Perriquet, 2004), Dynalign (Mathews and Turner, 2002), FoldalignM (Torarinsson, *et al.*, 2007), RNAalifold (Hofacker, *et al.*, 2002) and Stemloc (Holmes, 2005).

The correlation coefficient (CC) defined by Matthews (1975) is used to evaluate the prediction accuracy, which is approximated as the geometric mean of the prediction sensitivity (SEN) and specificity (SPE) (Gorodkin, *et al.*, 2001):

$$CC=\sqrt{SEN\cdot SPE}\ ,\quad 0\le CC\le 1$$

where SEN is the fraction of true base pairs that are predicted correctly, and SPE is the fraction of predicted base pairs that are true.

3.1 Prediction of common structures

We tested RNA Sampler and other programs on real sequences from 10 RNA regulatory motif or ncRNA gene families retrieved from the Rfam database (Griffiths-Jones, et al., 2005), including Cobalamin, gcvT, glmS, purine, RFN, sbox, THI, tRNA, U1 and yybP-ykoY. These data sets have been used in previous studies (Alkan, et al., 2006; Bafna, et al., 2006). Detailed information of all data sets is available in Supplementary Table 1. For each RNA family, Rfam provides a hand-curated seed alignment and a corresponding common structure, which are used to determine the exact base pairs in individual sequences.

First, we tested RNA Sampler on two-sequence sets generated from nearly all pairs of unique sequences in the seed alignment of each RNA family. RNA Sampler gave an aver-age *CC* of 0.60, ranging from 0.42 to 0.82 (Detailed values are listed in Supplementary Table 2). This was very close to the aver-age *CC* of 0.61 by Dynalign, which performed the best among all programs. However, RNA Sampler was significantly faster than Dynalign. The performance of Stemloc and RNAalifold were comparable to RNA Sampler on some RNA families, but worse on the others. RNA Sampler, Dynalign, Stemloc and RNAalifold all gave better *CC* than CARNAC. Overall, RNA Sampler gave a near best performance with a fast speed among all tested programs.

We then tested RNA Sampler and other programs on multiple-sequence sets of the 10 RNA families. To avoid sequence selection bias, we generated 100 sequence sets for each RNA family by randomly selecting 5 unique sequences from the Rfam seed alignment. These sequence sets cover nearly all unique sequences in the seed alignment, except for the tRNA family due to the large number of unique sequences in this family.

The prediction accuracy of RNA Sampler is significantly improved on multiple sequences compared to two sequences. Fig. 4 shows the *CC*, *SEN* and *SPE* of the structure predictions by RNA Sampler and other programs on the multiple-sequence sets, and detailed values are listed in Supplementary Table 3. Overall, RNA Sampler gave very good predictions on the 10 RNA families, with the average *CC*, *SEN* and *SPE* of 0.72, 0.73 and 0.72, respectively, ranging from 0.57 to 0.95. RNA Sampler performed the best among all tested programs, with the highest *CC* for 9 out of the 10 RNA families, and near highest for the other family.

RNA Sampler outperformed CARNAC and Stemloc on all tested RNA families. CARNAC only predicted partially correct structures, which led to poor sensitivity and relatively good specificity. The low performance of Stemloc might be partially due to the low *-nf* option used in the test (*-nf = 100*) (Holmes and Rubin, 2002; Holmes, 2005). Higher *-nf* settings were tried, however the corresponding higher memory requirement made the program very slow and often crash. RNAalifold requires reliable alignments for decent structure predictions. We used ClustalW alignments as inputs for RNAalifold in this study. RNAalifold performed the best on the RFN family sets that have high sequence identities, but had poor performance on the other sequence sets with low identities, in which ClustalW alignments might not be reliable. RNA Sampler performed as well as RNAalifold on the RFN family and gave significantly better results than RNAalifold on the other RNA families. Dynalign gave similar or slightly better sensitivities than RNA Sampler on the gcvT, sbox, yybp-ykoY and U1 families, but RNA Sampler showed much better sensitivities and overall performance on the other families. Besides, RNA Sampler always ran significantly faster than Dynalign. The good sensitivities of Dynalign can be partially explained by the fact that Dynalign cannot predict common structures on multiple-sequence sets and it was evaluated by the average performance on all pairwise predictions. Pairwise predictions give structures common in two sequences but not necessarily conserved in other sequences, which can lead to high sensitivities but poor specificities. FoldalignM had the second best overall performance among all programs. It showed as good sensitivities as RNA Sampler on the sbox, THI, tRNA and yybp-ykoY families, but lower sensitivities on the other families. RNA Sampler also ran much faster than FoldalignM (see Supplementary Fig. 4 and Supplementary Table 5).

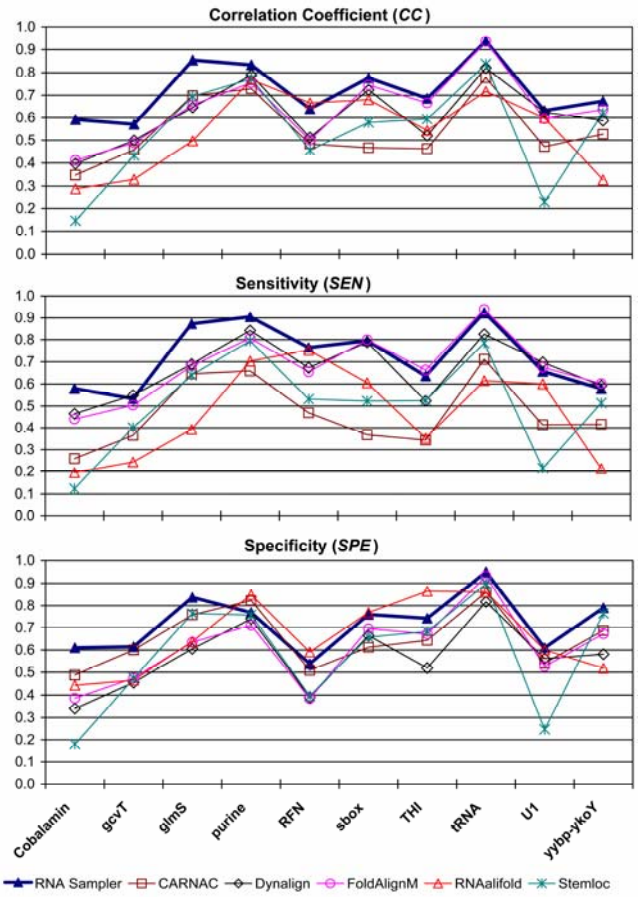

Fig. 4. Average correlation coefficient (*CC*), sensitivity (*SEN*) and specificity (*SPE*) of the common secondary structure predictions by RNA Sampler and other programs on multiple-sequence sets of 10 RNA families. Parameters used for RNA Sampler: *X = 0*, *sl = 3*, *d = 15*, *r = 15*, and *S = 75*. Stemloc, Dynalin, RNAalifold and CARNAC were all run with default parameters. For Stemloc *–nf = 100*; for Dynalign, *M = 15*; for RNAalifold, ClustalW alignments were used as input.

We also evaluated the prediction accuracy at the stem level (Bafna*, et al.*, 2006), where sensitivity (*SEN*) is the fraction of true stems that overlap with predicted stems, and specificity (*SPE*) is the fraction of predicted stems that overlap with true stems. With this measurement, the average *CC, SEN* and *SPE* of RNA Sampler on the 10 RNA families increased from 0.72, 0.73 and 0.72 at the base pair level to 0.77, 0.80 and 0.76, respectively (see Supple-

mentary Fig. 2 and Supplementary Table 4). Other programs also showed increased *CCs* at the stem level, and the relative performance between RNA Sampler and other programs are similar to that at the base pair level.

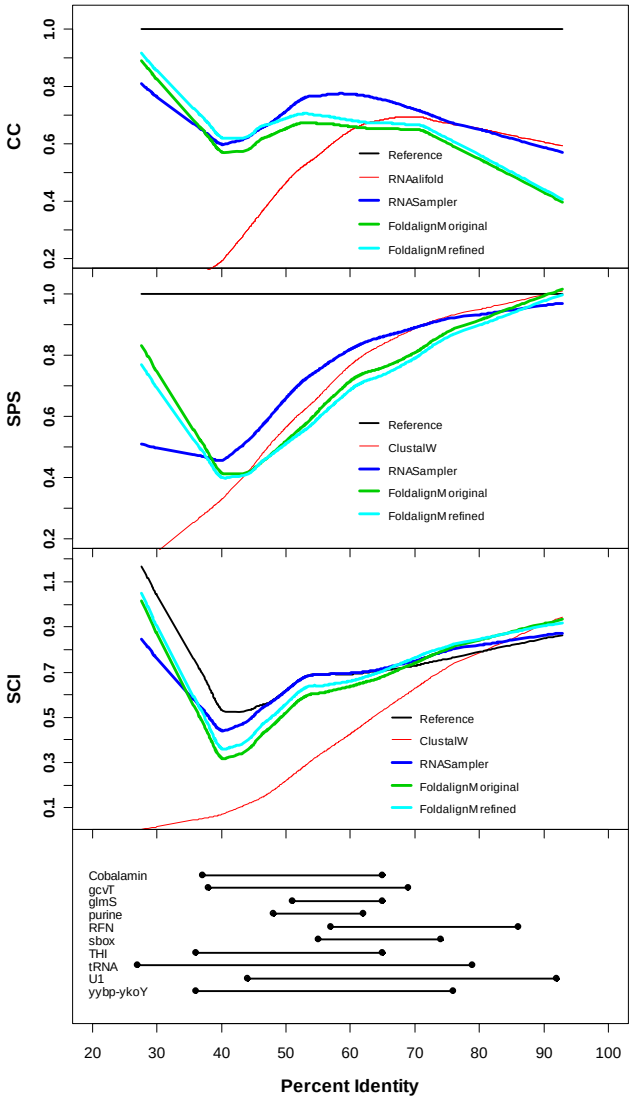

Fig. 5. Comparison of average *CC*, *SPS* and *SCI* among RNA Sampler, FoldalignM and RNAalifold (on ClustalW alignments) on sequences of a wide range of identities from 10 RNA families. The Rfam seed alignments and structures were used as benchmarking references. Only reliable alignments in the stem regions of the Rfam seed alignments were examined in calculating the *SPS* scores. The curves were generated using lowess (locally weighted regression) smoothing. The black bars below the graph represent the ranges of sequence identities for the 10 RNA families.

3.2 Prediction of structural alignments

We employed and extended the method proposed in a previous study on two-sequence structural alignment (Gardner, et al., 2005) to evaluate the performance of RNA Sampler on multiple-sequence structural alignment across a wide range of sequence identities. We used the sum-of-pairs score (*SPS*) (Thompson, et al., 1999), the fraction of aligned base pairs in the reference alignment that are correctly aligned in the predicted alignment, to measure the accuracy of structural alignments at the base pair level. The structural conservation index (*SCI*) calculated by RNAz (Hofacker, et al., 2004) was used to measure the conserved structure information contained in the predicted alignment. A *SCI* score ranges from 0 to ~1. If the consensus structure is supported by compensatory or consistent mutations, *SCI* can be greater than 1.

For each of the 10 RNA families, we randomly regenerated multiple sequence sets that evenly cover a wide range of mean pairwise sequence identities. Each family has a certain range of sequence identities (shown in Fig. 5), which is generally between 40 % and 80% but can be lower than 40% for the tRNA family and higher than 80% for the U1 family. Using the Rfam seed alignments and common structures as benchmarking references, we compared the alignments and structures predicted by RNA Sampler and FoldalignM with those by RNAalifold, which took sequence-based ClustalW alignments as inputs. We plotted *CC*, *SPS* (only on the reference stem regions) and *SCI* as functions of the mean pairwise sequence identity on each RNA family (see Supplementary Fig. 7). The shapes of the plots vary between RNA families but with a similar trend. Fig. 5 shows the average *CC*, *SPS* and *SCI* on all the 10 RNA families combined. It should be noted that the plot below 40% and above 80% sequence identities only represent the tRNA and U1 family, respectively.

The *CC* plot in Fig. 5 shows that RNA Sampler gives overall better structure prediction than FoldalignM by 10 ~ 20% across the whole sequence identity range except for below 40% on the tRNA data sets. RNA Sampler gives significantly better predictions than RNAalifold across the whole identity range below 80%, especially in the low identity region, and gives as good predictions as RNAalifold on sequence sets of identities above 80%, the U1 sets.

The *SPS* curves of RNA Sampler, FoldalignM and ClustalW show a similar trend: the *SPS* scores tend to decrease with the decrease of sequence identities. RNA Sampler gives better *SPS* scores than FoldalignM and ClustalW across a wide sequence identity range within which most of the RNA families are located, 40% ~ 70%. This indicates that RNA Sampler gives overall better structural alignments. RNAalifold's structure prediction performance (*CC*) is directly correlated with the quality of its input ClustalW alignment (indicated by *SPS*), which deteriorates as the sequence identity decreases. RNA Sampler gives lower *SPS* scores than FoldalignM in the low identity range (<40%), partially due to the relatively less accurate structure predictions on some tRNA sets in that identity range. In addition, the Rfam alignments might be imperfect in some cases, which could lead to an underestimate of the alignment accuracy of RNA Sampler. Indeed, we observed some examples in which perfectly matched consecutive base pairs were misaligned in the stem regions of the Rfam and FoldalignM alignments but correctly aligned by RNA Sampler, however, in such cases the RNA Sampler alignments were classified as wrong since the Rfam alignments were used as benchmarks. FoldalignM reports two structure and alignment outputs: one is original and the other is refined. Interestingly, the original outputs of FoldalignM are often slightly better aligned (reflected by higher *SPS* scores) but less accurate in terms of base pairings (with lower *CC*) than the refined outputs. This discrepancy implies that slightly different structural alignments can correspond to the same common struc-

ture, and that the accuracy of structural alignments might be underestimated by using *SPS* scores alone.

Although the structural alignments by RNA Sampler and FoldalignM do not completely match the Rfam reference alignments, they do give high *SCI* scores approaching those of the reference alignments across the whole identity range, indicating that their alignments have indeed caught the conserved structure information. RNA Sampler does better than FoldalignM with higher *SCI* scores in the 40 ~ 70% identity range. In contrast, the ClustalW alignments have significantly lower *SCI* scores across almost the whole identity range except the very high identity region (>70%), suggesting that these sequence-based alignments could not preserve the conserved structure information.

Overall, RNA Sampler not only predicts common structures more accurately than other programs, but also provides better structural alignments that more accurately represent the RNA secondary structure conservations, across a wide range of sequence identities.

3.3 Prediction of pseudoknot structures

Because our sampling strategy has no limitations on how compatible blocks can be assembled into structures, RNA Sampler can predict pseudoknot structures if the user allows it to sample blocks forming crossovers. We introduced a user definable parameter, X, the maximum number of crossovers allowed in a structure. By default, no crossovers are allowed ($X = 0$) for more accurate predictions on non-pseudoknot structures, and the ($X > 0$) option is recommended only if the user suspects that there may be pseudoknots present in the structures.

Table 1. Performance of RNA Sampler on multiple-sequence sets containing pseudoknot structures.

	CC[a]	*SEN*[a]	*SPE*[a]	correct / total stems	Runtime (s)
10 α operon mRNA leaders[b]	0.38	0.46	0.31	2/4	253
8 S15 mRNA leaders (pseudoknot)	0.76	0.74	0.79	3/3	662
8 S15 mRNA leaders (stem-loop)	0.66	0.67	0.66	3/3	656

a. *CC*, *SEN* and *SPE* are calculated based on matching of base pairs between predicted structures and known structures in the *E. coli* sequence only.

b. Parameters used for RNA Sampler: $r = 15$, $S = 100$, $sl = 4$, $d = 15$; $X = 2$ for the S15 set allowing pseudoknots, and $X = 0$ for the S15 set not allowing pseudoknots.

We tested RNA Sampler on a few sequence sets with known pseudoknot structures (see Table 1), including a set of 10 bacterial α operon mRNA leader sequences and a set of 8 bacterial S15 mRNA leader sequences used in a previous study (Ji, *et al.*, 2004). RNA Sampler only correctly predicted 46% of the known base pairs on the α operon set, but two stems forming the core pseudoknot structure were correctly predicted (see Supplementary Fig. 3A). One of the predicted stems is not in the published α operon RNA structure, but it was also predicted by comRNA in the previous study (Ji, *et al.*, 2004). The S15 leader sequences may form two alternative structures: one contains a pseudoknot, and the other is a non-pseudoknot stem-loop structure. At the base pair level, RNA Sampler gave sensitivities of 0.74 and 0.67 on the two alternative structures, respectively; it missed a few base pairs in the long stems, as no bulge or loop was allowed in stems. However, at the stem level, RNA Sampler correctly predicted all 3 stems in the pseudoknot structure, and all 3 stems in the alternative stem-loop structure if crossovers were not allowed (see Supplementary Fig. 3B). The performance of RNA Sampler on these data sets is slightly worse than that of comRNA but better than most of the other programs tested in the comRNA study (Ji, *et al.*, 2004). The runtimes of RNA Sampler on these data sets were 4 ~ 10 minutes.

3.4 Complexity of the algorithm

The speed of the algorithm is limited by the iteration step. In each iteration, it takes $O(m \cdot n)$ time to generate $m \cdot n$ possible blocks by comparing all m stems in sequence A with all n stems in sequence B. The introduction of the parameter d, the maximum shift distance allowed in a block between two stems, and the constraint of picking only stem pairs with reciprocal best conversation scores as blocks reduce the total number of blocks for sampling from $m \cdot n$ to m (assume that $m \leq n$). The sampling procedure takes $O(m)$ time to eliminate conflicting blocks by comparing the blocks with the chosen one. Therefore, the time complexity for predicting the common structure between two sequences is $O(m^2 \cdot r \cdot S)$, where r is the total number of iterations, and S is sample size, *i.e.* number of structures sampled in each iteration. For a multiple-sequence set, RNA Sampler samples common structures between all pairs of sequences, and thus the time complexity becomes $O(m^2 \cdot r \cdot S \cdot N^2)$, where N is the number of sequences.

Supplementary Table 5 and Supplementary Fig. 4 list the runtimes of RNA Sampler and other programs on the 10 RNA families on a Linux machine with a Pentium IV 3.0 GHz CPU and 3 GB RAM. On a data set containing 5 sequences whose average length ranges from 70 to 200 nt, RNA Sampler took 6 ~ 180 s to complete the structure prediction. Data sets of longer sequences contain more stems and would take more iterations to converge and thus longer runtime. CARNAC, RNAalifold and Stemloc (*-nf* = 100) ran faster or at a similar speed compared to RNA Sampler, but RNA Sampler was much more accurate; RNA Sampler ran much faster than Dynalign and FoldalignM and also gave better prediction accuracy.

4 DISCUSSION

We developed a new sampling based algorithm for common RNA secondary structure prediction and structural alignment on multiple sequences. In extensive testing on various RNA families, our algorithm showed very good prediction accuracy and speed, giving better performance than other existing programs for common RNA secondary structure prediction. Besides, our algorithm is able to predict pseudoknot structures.

The consistency-based method significantly improves the performance of RNA Sampler on multiple-sequence sets compared to its performance on two-sequence sets. It effectively incorporates structure information from all pairwise sequence comparisons into the consensus structure of multiple sequences. The stems shared by multiple sequences will reinforce each other to be sampled more frequently in subsequent iterations, while non-conserved stems will fade away. This approach has also been successfully applied to the multiple sequence alignment problem (Do, *et al.*, 2005). Whereas other programs, such as Stemloc (Holmes and Rubin, 2002; Holmes, 2005) and Foldalign (Gorodkin, *et al.*, 2001), usually apply the progressive method, which picks seed consensus struc-

tures based on pairwise sequence comparisons and progressively adds more sequences to the structure. The progressive method is dependent on the order of seed structure picking, and wrongly picked seed structures could lead to wrong common structures; whereas the consistency-based method integrates all information available to reach consensus.

RNA Sampler runs quickly on medium sized data sets (see Supplementary Table 5). Our data show that the sampled structures converge quickly usually after a few iterations (see Supplementary Fig. 1). Since the consistency-based method for structure prediction on multiple sequences takes consideration of all pairwise sequence comparisons in each iteration, it is computationally more expensive than the progressive method. To reduce runtime, we designed a fast comparison approach (see Supplementary Appendix IV): instead of doing all pairwise sequence comparisons, it randomly picks one sequence to compare with all other sequences in each iteration. With the fast approach, the runtimes of RNA Sampler on the 10 RNA families were reduced by about 2-fold and the prediction accuracies were slightly worse than those of the slow approach but still better than those of other programs overall (see Supplementary Fig. 5 and Supplementary Table 6). The fast comparison approach provides an option for users to quickly screen a large number of sequence sets for presence of common structures.

The performance of a sampling algorithm is usually affected by the sample size. A large enough sample size is needed for robust structure prediction. We tested our algorithm on the five-sequence data sets for all the 10 RNA families with different sample sizes, at 100, 200 and 300. We found that these different sample sizes did not make much difference on the prediction results, and that sampling structures 100 times in each iteration was sufficient for making robust predictions. The constraint that a block has to be reciprocal best stem pair dramatically reduces the number of blocks to be sampled and alleviates the need for a huge sample size. This constraint causes little loss of accuracy in the predictions.

Currently we only consider perfectly aligned stem pairs in the structure sampling procedure, and the accuracy might be improved by allowing bulged stems. Estimating initial base pairing probabilities with pseudoknots is not trivial, and we currently use a fast heuristic method (see Supplementary Appendix II). Dirks and Pierces (2004) proposed a partition function based algorithm to calculate base pairing probabilities allowing pseudoknots, but the algorithm is computationally expensive, with a time complexity of $O(L^5)$, where L is the length of the sequence. Implementing this algorithm into the initialization step will possibly improve the performance of RNA Sampler in predicting pseudoknot structures but require extra runtime.

Although overall our algorithm performs well (*CC* ranges between 0.57 and 0.94) and performs better than other existing programs, in many cases it only predicts partially correct structures. The heuristic sampling procedure and the unique structure conservation measuring function introduced to the algorithm make it run fast and meanwhile achieve good prediction accuracy, which indicates the effectiveness of the algorithm for solving the challenging problem of common RNA structure prediction. On the other hand, the heuristic procedure compromises the optimal performance of the algorithm. Besides common structure prediction, an important output of RNA Sampler is the structural alignment, which better represents the RNA secondary structure conservation information than the sequence alignment generated by algorithms such as ClustalW. The structural alignment output can be used as input for the RNA structure detection programs, such as RNAz (Washietl*, et al.*, 2005). Our preliminary tests have shown that the structural alignments generated by RNA Sampler can improve the sensitivities of RNAz in detecting the presence of RNA motifs in sequences of low identities.

ACKNOWLEDGEMENTS

This work is supported by the NIH grant HG00249 and the DOE grant ER63972. We thank the authors of all software used or tested in this study for making their software publicly available.

REFERENCES

Alkan, C., Karakoc, E., Sahinalp, C., Unrau, P., Ebhardt, A., Zhang, K. and Buhler, J. (2006) RNA Secondary Structure Prediction via Energy Density Minimization. Research in Computational Molecular Biology (RECOMB). Venice, Italy.

Bafna, V., Tang, H. and Zhang, S. (2006) Consensus folding of unaligned RNA sequences revisited, J Comput Biol, 13, 283-295.

Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function, Cell, 116, 281-297.

Cary, R.B. and Stormo, G.D. (1995) Graph-theoretic approach to RNA modeling using comparative data, Proc Int Conf Intell Syst Mol Biol, 3, 75-80.

Dirks, R.M. and Pierce, N.A. (2004) An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots, J Comput Chem, 25, 1295-1304.

Do, C.B., Mahabhashyam, M.S., Brudno, M. and Batzoglou, S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment, Genome Res, 15, 330-340.

Do, C.B., Woods, D.A. and Batzoglou, S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models, Bioinformatics, 22, e90-98.

Eddy, S.R. (2001) Non-coding RNA genes and the modern RNA world, Nat Rev Genet, 2, 919-929.

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

Gardner, P.P., Wilm, A. and Washietl, S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs, Nucleic Acids Res, 33, 2433-2439.

Giegerich, R., Voss, B. and Rehmsmeier, M. (2004) Abstract shapes of RNA, Nucleic Acids Res, 32, 4843-4851.

Gorodkin, J., Stricklin, S.L. and Stormo, G.D. (2001) Discovering common stem-loop motifs in unaligned RNA sequences, Nucleic Acids Res, 29, 2135-2144.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes, Nucleic Acids Res, 33, D121-124.

Gutell, R.R., Power, A., Hertz, G.Z., Putz, E.J. and Stormo, G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods, Nucleic Acids Res, 20, 5785-5795.

Hofacker, I.L., Bernhart, S.H. and Stadler, P.F. (2004) Alignment of RNA base pairing probability matrices, Bioinformatics, 20, 2222-2227.

Hofacker, I.L., Fekete, M. and Stadler, P.F. (2002) Secondary structure prediction for aligned RNA sequences, J Mol Biol, 319, 1059-1066.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M. and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures, Monatsh Chemie, 125, 167-188.

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

Holmes, I. and Rubin, G.M. (2002) Pairwise RNA structure comparison with stochastic context-free grammars, Pac Symp Biocomput, 163-174.

Ji, Y., Xu, X. and Stormo, G.D. (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences, Bioinformatics, 20, 1591-1602.

Knudsen, B. and Hein, J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history, Bioinformatics, 15, 446-454.

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

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.

Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim Biophys Acta, 405, 442-451.

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

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

Muckstein, U., Hofacker, I.L. and Stadler, P.F. (2002) Stochastic pairwise alignments, Bioinformatics, 18 Suppl 2, S153-160.

Rivas, E. and Eddy, S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots, J Mol Biol, 285, 2053-2068.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R.C. and Haussler, D. (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.

Siebert, S. and Backofen, R. (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons, Bioinformatics, 21, 3352-3359.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. and Steinberg, S. (1998) Compilation of tRNA sequences and sequences of tRNA genes, Nucleic Acids Res, 26, 148-153.

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J. and Giegerich, R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes, Bioinformatics, 22, 500-503.

Stormo, G.D. and Ji, Y. (2001) Do mRNAs act as direct sensors of small molecules to control their expression?, Proc Natl Acad Sci U S A, 98, 9465-9467.

Tabaska, J.E., Cary, R.B., Gabow, H.N. and Stormo, G.D. (1998) An RNA folding method capable of identifying pseudoknots and base triples, Bioinformatics, 14, 691-699.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res, 22, 4673-4680.

Thompson, J.D., Plewniak, F. and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs, Nucleic Acids Res, 27, 2682-2690.

Torarinsson, E., Havgaard, J.H. and Gorodkin, J. (2007) Multiple structural alignment and clustering of RNA sequences, Bioinformatics.

Touzet, H. and Perriquet, O. (2004) CARNAC: folding families of related RNAs, Nucleic Acids Res, 32, W142-145.

Washietl, S., Hofacker, I.L. and Stadler, P.F. (2005) Fast and reliable prediction of noncoding RNAs, Proc Natl Acad Sci U S A, 102, 2454-2459.

Winkler, W.C. (2005) Riboswitches and the role of noncoding RNAs in bacterial metabolic control, Curr Opin Chem Biol, 9, 594-602.

Zuker, M. (1994) Prediction of RNA secondary structure by energy minimization, Methods Mol Biol, 25, 267-294.