GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Novel representation of RNA secondary structure used to improve prediction algorithms

Q. Zou[1], C. Lin[1], X.-Y. Liu[2], Y.-P. Han[3], W.-B. Li[3] and M.-Z. Guo[2]

[1]School of Information Science and Technology, Xiamen University, Xiamen, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[3]Key Laboratory of Soybean Biology, Chinese Education Ministry, Soybean Research Institute, Northeast Agricultural University, Harbin, China

Corresponding author: M.-Z. Guo
E-mail: maozuguo@hit.edu.cn

Genet. Mol. Res. 10 (3): 1986-1998 (2011)
Received November 29, 2010
Accepted July 24, 2011
Published September 9, 2011
DOI http://dx.doi.org/10.4238/vol10-3gmr1181

ABSTRACT. We propose a novel representation of RNA secondary structure for a quick comparison of different structures. Secondary structure was viewed as a set of stems and each stem was represented by two values according to its position. Using this representation, we improved the comparative sequence analysis method results and the minimum free-energy model. In the comparative sequence analysis method, a novel algorithm independent of multiple sequence alignment was developed to improve performance. When dealing with a single-RNA sequence, the minimum free-energy model is improved by combining it with RNA class information. Secondary structure prediction experiments were done on tRNA and RNAse P RNA; sensitivity and specificity were both improved. Furthermore, software programs were developed for non-commercial use.

Key words: RNA secondary structure; Dot plots; Stem; Comparative sequence analysis

INTRODUCTION

During the last few years, many studies have shown the key function of non-coding RNA (ncRNA). For instance, microRNAs play important roles in the post-transcriptional mechanism, and small nucleolar RNAs (snoRNAs) can guide RNA ribose methylation and pseudouridylation. Secondary structure information is necessary when mining these non-coding genes, analyzing the function of ncRNA or predicting RNA-RNA interaction.

There are two main RNA secondary structure prediction approaches according to different input data. For a single sequence, the minimum free-energy model is the best choice, but it is not as accurate as the comparative sequence analysis (CSA) method (Gardner and Giegerich, 2004), which suits dealing with a group of homologous sequences. However, multiple sequence alignment (MSA) is always used in CSA and may influence accuracy if the homologous sequences are not similar. Furthermore, RNA class information may be helpful for secondary structure prediction. When non-coding RNA genes are mined or RNA-RNA interactions are predicted, the class information is always known by biology scientists. Thus, it is meaningful to get more accurate structure using the class information. In this paper, we propose a novel representation of RNA secondary structure and a convenient structure comparison method. We then try to solve the above-mentioned problems.

MATERIAL AND METHODS

A novel representation of RNA secondary structure

Comparing different RNA secondary structures is time-consuming if structure is represented as a forest (Zhang et al., 2007). We consider that secondary structure consists of stems and that similar secondary structures always have similar stems. If two stems in different RNAs are similar, it means that the length and the position of the two stems are similar, as shown in Figure 1.

From Figure 1, we can see that if we describe the position of a stem with two center points, the points will fall in small regions when stems are homologous in different RNA sequences. Thus, we can compare the position of different stems with the distances of the center points and describe a stem position with information on the two center points' positions.

Figure 2 gives a geometric expression of our concept. When facing bugles or internal loops, the concept can also be extended as the average of the first and the last base pair position.

Bugles and internal loops can be viewed as broken lines in dot plots. The center is shifted as the line is broken, as shown in Figure 3.

When predicting secondary structure, candidate stems are always listed in a "stem pool" first. Then, position relation is computed for every two stems in the stem pool. The position relation can be categorized as compatible, conflicting and pseudoknot. If a stem is drawn in dot plots, the regions of different position relation can be easily identified as in Figure 4. Therefore, we can easily calculate the relationship of two different stem candidates in one sequence by the centers and length information.

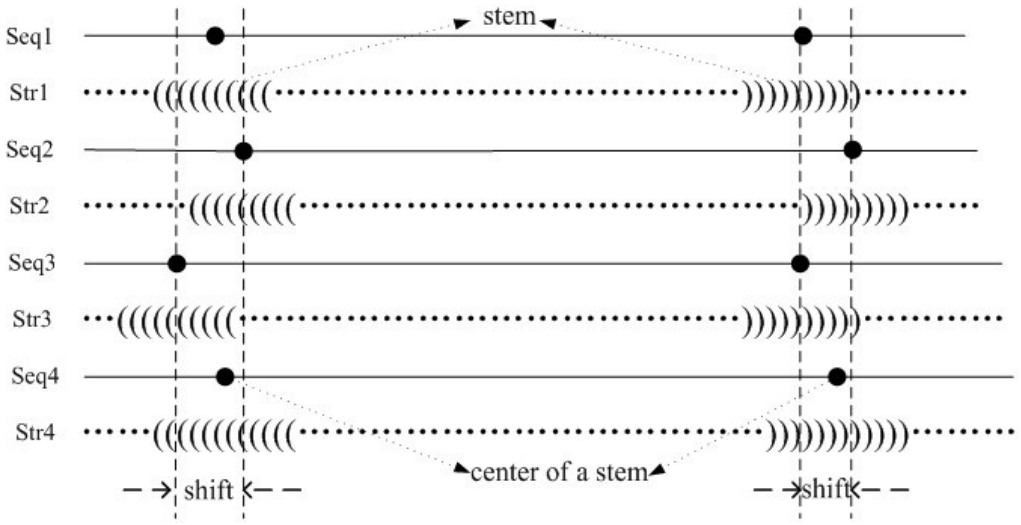

Figure 1. Position-conservative stems in homologous RNA sequences.

Definition 1. For a possible stem $((a_i,a_{i+1},...,a_{i+k}), (a_{j-k},a_{j-k+1},...,a_j))$ in an RNA sequence $\{a_1,a_2,...,a_n\}$, we define $(i+k/2,j-k/2)$ to be the center of this stem. Dot plots are always used in bioinformatics (Leland, 1999). When representing RNA secondary structure, the stem is a diagonal line in dot plots and the center of a stem is the coordinate values of the diagonal's center, as shown in Figure 2.

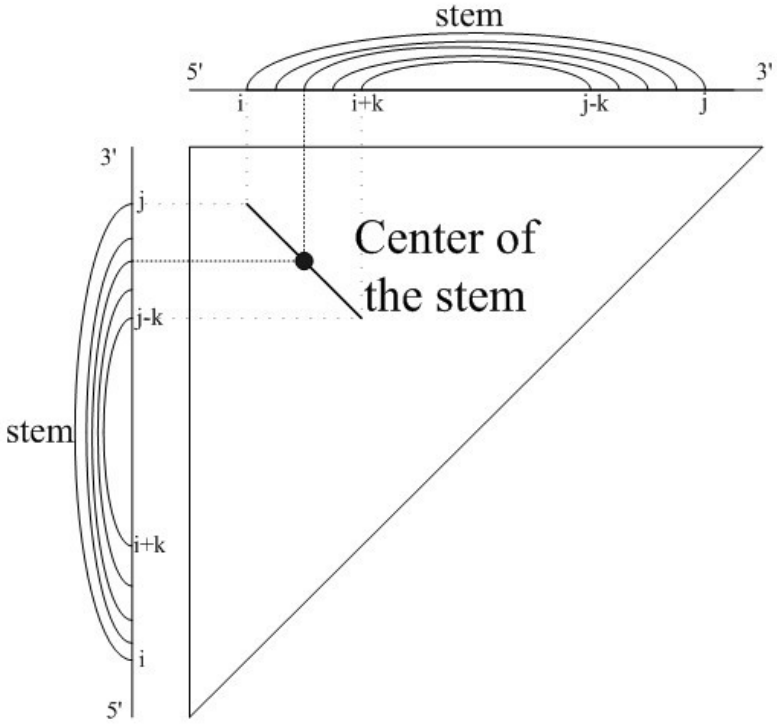

Figure 2. The geometric meaning of center of stem.

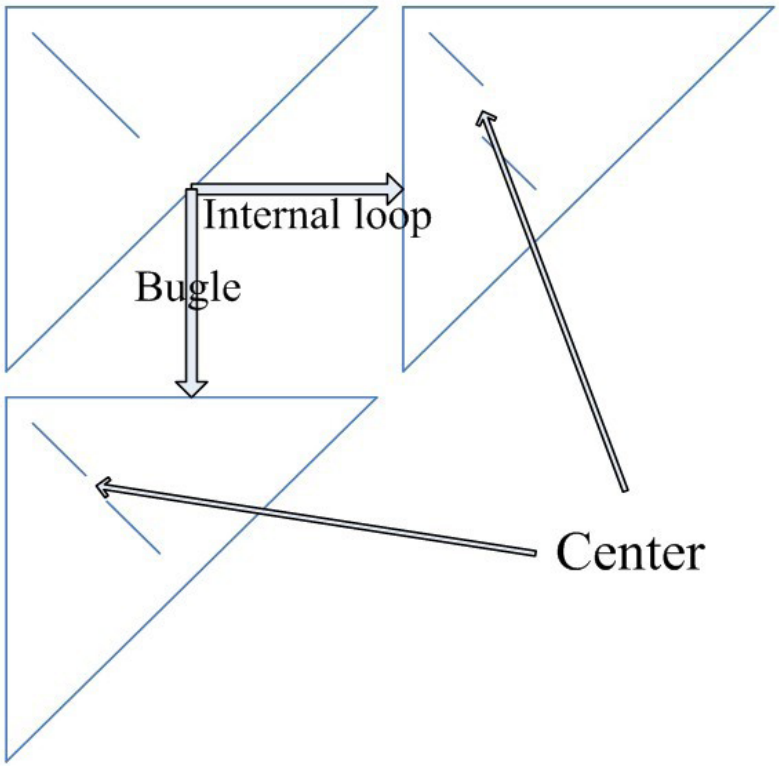

Figure 3. Centers of a bugle and internal loop.

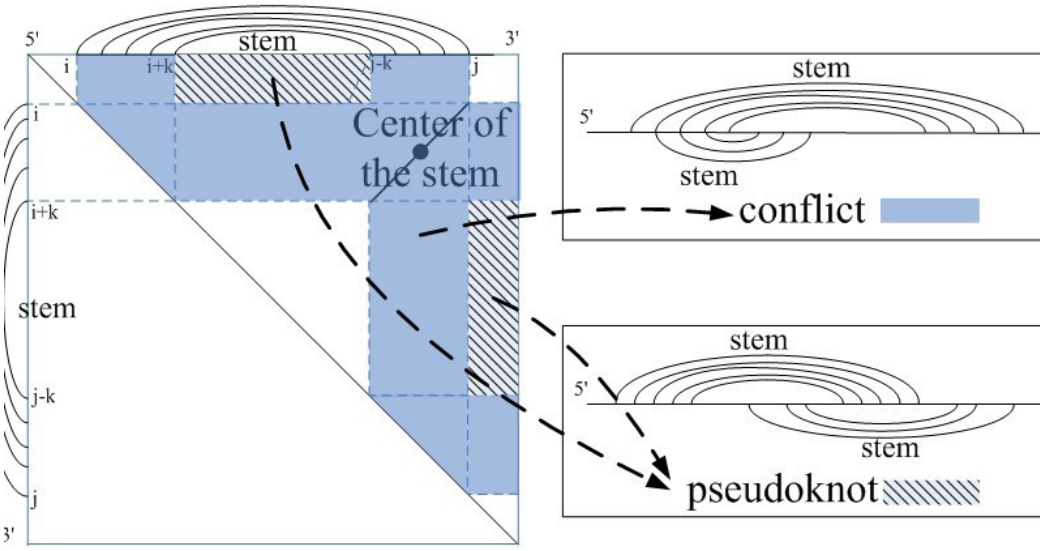

Figure 4. Position relation in dot plots.

The center only tells the position information of a stem. A stem can be fixed with the center together with length information. Therefore, if we want to assess the difference between two stems in two different RNA sequences, we should consider both the distance between centers and the length distance.

First, we discuss the distance between centers. For different homologous RNA sequences, the length is always different. Thus, alignment will be useful when assessing the position difference of two stems. After alignment, the center can be viewed as the center point of the stem in the aligned sequence. The distance between the aligned centers together with the length difference can then be used to assess the difference between two stems, as shown in Figure 5.

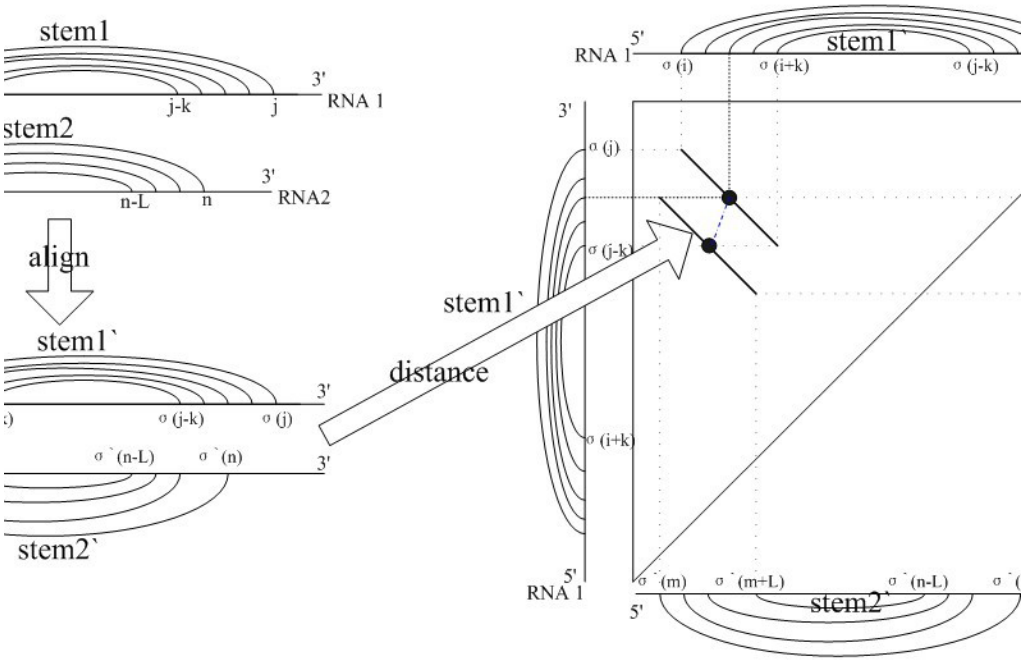

Figure 5. Distance between two stems.

Definition 2. Given homologous RNA sequences, for any two stems from different RNAs, we denote the distance between the two stems as $D(S_1,S_2) = dis(S_1,S_2)+\lambda*|L_1-L_2|$ in which S_1 and S_2 are two stems from different homologous RNA sequences, L_1 and L_2 are their lengths, $dis(S1,S2)$ is the Euclidean distance between the centers in the aligned sequences, and λ is a weight parameter. Based on these concepts, we can easily assess the position of stems. Furthermore, different stems together with secondary structure can be compared quickly. Therefore, we can improve the two main secondary structure prediction approaches using our novel representation.

All current multiple sequence alignment algorithms attempt to search the most similar alignment. However, incorrect gaps or spaces may disconnect real stems. If gaps are inserted incorrectly into a stem, it is difficult to predict it as a whole stem with the prediction software.

Comparative sequence analysis method independent of multiple sequence alignment

Comparative sequence analysis methods are effective in RNA secondary structure prediction, when dealing with a group of homologous sequences with consensus structure.

Most of the comparative sequence analysis methods are based on multiple sequence alignment. They attempt to find the consensus structure or covariance information from the aligned sequences, as shown in Figure 6. However, structural information may be ignored in the first step, and incorrect gaps or spaces may be involved when doing multiple sequence alignment.

From Figure 7, we can see that the incorrect gaps in the 22nd to 27th sequences will make the corresponding bases (70 and 69) unpaired. Thus, the acceptor arm is improperly folded, as shown in Figure 7. 76 tRNA sequences are aligned with ClustalW as shown in the upper part of this figure. They are *Acinetobacter* sp ADP1 from the Genomic tRNA Database (Lowe and Eddy, 1997). The left secondary structure is the prediction of the 22nd sequence from Pfold (Knudsen and Hein, 1999), while the right is the real structure drawn by PseudoViewer3 (Byun and Han, 2006). The real acceptor should be a 7-base length stem, but Pfold gives an internal loop.

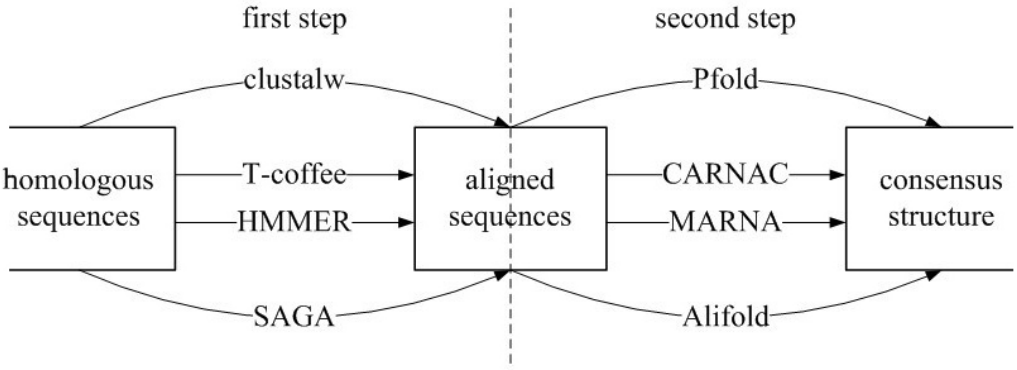

Figure 6. The process of comparative sequence analysis method.

Here we propose a novel algorithm for predicting secondary structure for homologous RNA sequences. Our algorithm takes advantage of the "center" concept mentioned above and avoids the influence of multiple sequence alignment. Next is the detailed description, and Figure 8 shows the algorithm flow chart (Zou et al., 2008).

If there are k sequences and the average length is n, it is easy to determine that all the 7 steps cost $O(kn^3)$ time and $O(kn^2)$ memory. The process will be repeated r times, where r is less than the number of the stem candidates in one sequence. Since $r<<n$ and $k<<n$, the overall complexity is $O(n^3)$ time and $O(n^2)$ memory, which is the same as with RNAalifold (Hofacker, 2003) and hxmatch (Witwer et al., 2004).

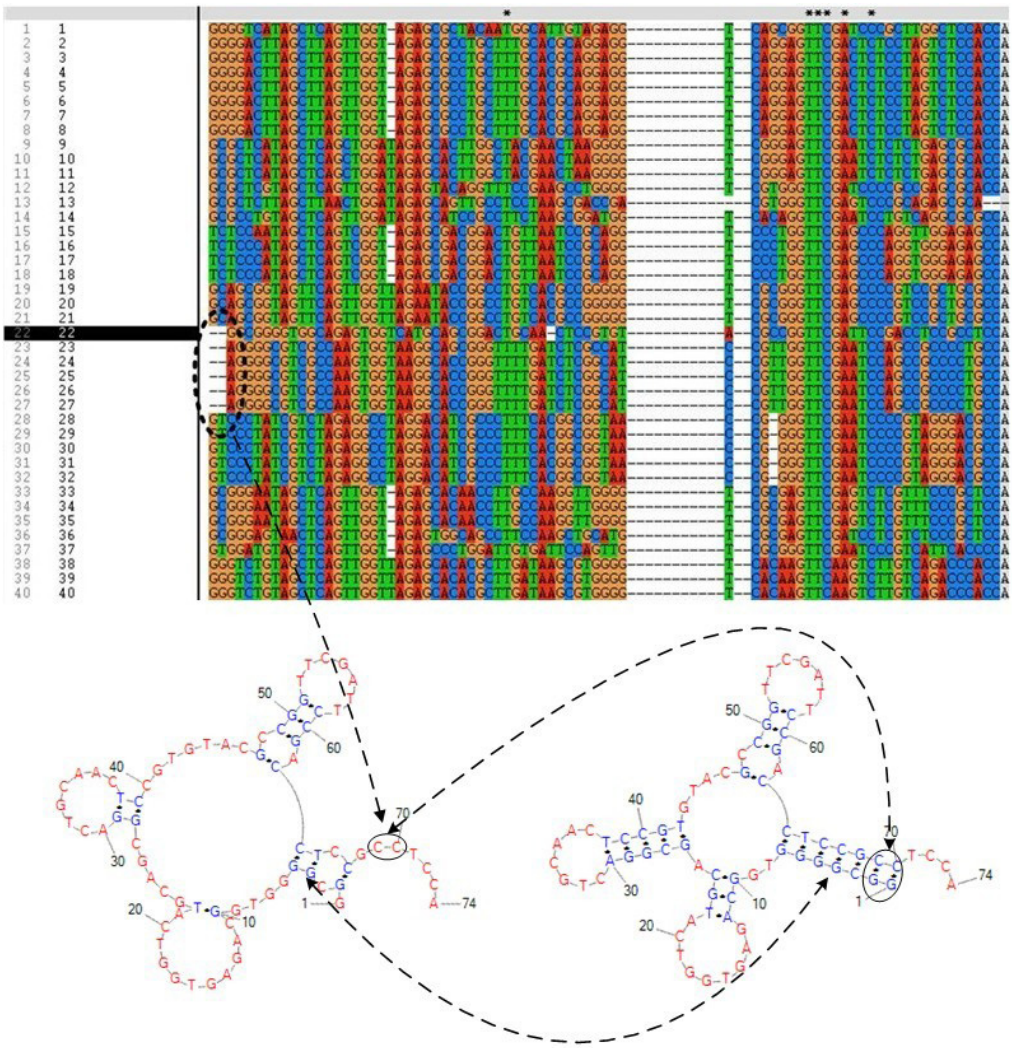

Figure 7. An example of multiple sequence analysis influence on secondary structure prediction.

Further supplement on minimum free-energy model

Most of the time, we do not have enough homologous sequences to fold. The minimum free-energy model is the best and the only choice for dealing with a single sequence. However, many special RNA structures are very different from the minimum free-energy ones. For instance, C/D box snoRNA is a short stem with a big loop. The sequence of this loop may fold into a more stable structure with lower energy. However, other molecules will interact with the C/D box snoRNA and influence the structure. Therefore, we can conclude that sequence information alone cannot decide the structure, further supplemental information is important when predicting secondary structure.

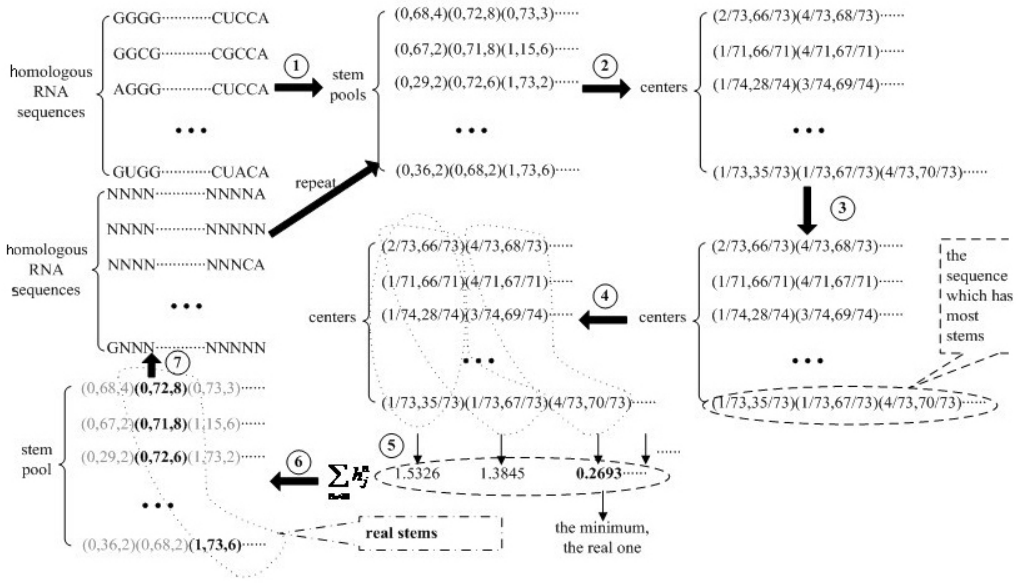

Figure 8. Algorithm flow chart of folding homologous RNA sequences independent of multiple sequence analysis.

Class information is useful, because biology knowledge can tell the approximate shape of the predicted RNA secondary structure. RNA sequences from same class always have similar secondary structure. If we know the class of a predicted RNA, we can fold it as a known structure of other sequence in this class. This idea is similar to "threading" in protein folding.

Class information is always known if the sequence is known. An unknown sequence can be aligned with an RNA database, such as RFAM (Griffiths-Jones et al., 2005) and ERPIN (Lambert et al., 2004). Even for a new class, which is not recorded in these databases, a new technique called atomic force microscopy (AFM) (Rounsevell et al., 2004) can also tell the approximate shape. Here, we fold the RNA sequence to an approximate shape using the concepts introduced above. If we know the structure together with a homologous sequence, we can first align the sequence with the predicted one, and then for every stem in the known structure, we search in the stem pool of the predicted ones for the one with minimum distance. If a homologous sequence is unknown and the approximate shape comes from AFM, we can also search the stems with Definition 2. Differently, we should denote the stems in approximate shape with Definition 1.

This is the main idea of our method. Stems are predicted by computing the minimum distance with the stems in approximate shape. However, the predicted stems sometimes conflict, especially when the conflicting stem is real but longer than the native stem. This problem was discussed in detail by Zou et al. (2009b), and a software was also developed, called *zqfold*.

RESULTS

We use tRNA to test the performance of our algorithms. Sensitivity and specificity are selected to measure and compare with other software. The structures of tRNA come from

Lowe and Eddy (1997) and Chan and Lowe (2009). This database divides tRNA into three kinds: Bacteria, Eukarya and Archaea. We tested the algorithm on three groups of tRNAs selected from *Halobacterium* sp, *Plasmodium falciparum* and *Anaplasma marginale* St. Maries.

First, we compare our method on a group of homologous sequences with Pfold (Knudsen and Hein 1999), MARNA (Siebert and Backofen, 2005) and CARNAC (Touzet and Perriquet, 2004). These three software programs come from three approaches of CSA method. The first approach is to do an MSA first, then to fold. Pfold is the main software program of this approach. For Pfold, homologous sequences are first aligned by ClustalX (Larkin et al., 2007) with default parameters. The second approach is to perform alignment and folding simultaneously. It is time-consuming, so CARNAC from this approach cannot deal with sequences longer than 80 nt. The third approach is to fold the sequences first, then to align all the structures to find a consensus structure. MARNA is an example software of this approach.

The performance is shown in Figure 9. We can see that our method performs as well as Pfold, and outperforms MARNA and CARNAC with regard to both sensitivity and specificity.

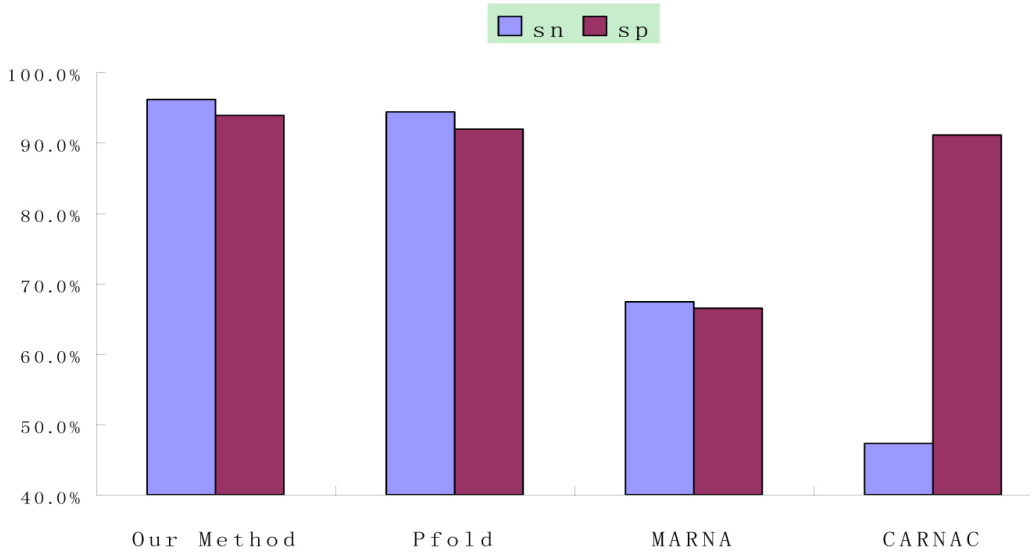

Figure 9. Performance comparison on multiple-sequence folding of tRNA. sn = TP / (TP + FN); sp = TP / (TP + FP).

Here, we can conclude that the first approach performs better than the other two approaches, and that tRNA is conservative for MSA, and therefore, Pfold can perform well. If more complex or less conservative sequences are folded, our method will work better.

Furthermore, we also tested our method on simple sequence folding. With regard to a tRNA molecule, our method performs better than the main software based on a minimum free-energy model, including RNAfold (Hofacker, 2003), mfold (Zuker, 2003) and Srna (Ding et al., 2004), as shown in Figure 10.

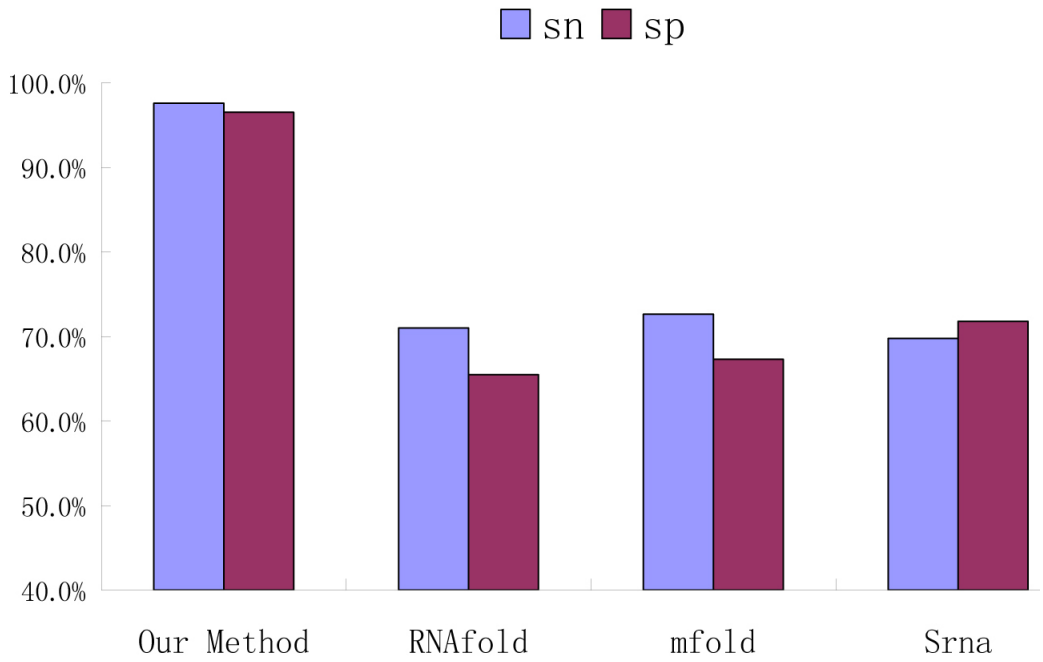

Figure 10. Performance comparison on single-sequence folding of tRNA. sn = TP / (TP + FN); sp = TP / (TP + FP).

It is known that tRNA is a class of simple and conservative molecules. We also tested RNAse P RNA, which is more complex and less conservative. Here, we study SM-A46(74), Lake Griffy B#41, Volunteer ESH212C, and Pond Scum#26 (Brown, 1999). The average performance with these four molecules also shows the robustness of our algorithms. Figures 11 and 12 show the performance comparison of multiple folding and single folding, respectively.

DISCUSSION

The above experiments prove the performance of our representation and prediction algorithms. As is known, secondary structure prediction is always a preliminary process for some problems, rather than an ultimate process. Successful application is much more meaningful than prediction accuracy. Here, we apply our prediction method instead of RNAfold in microRNA and snoRNA mining research, and achieve improvement.

First, we apply our prediction algorithms on mining H/ACA box snoRNA. H/ACA box snoRNA is a class of non-coding RNA, which is usually required for regulating RNA pseudouridylation. It has a hairpin-hinge-hairpin-tail consensus structure, as shown in Figure 13. Secondary structure features are selected when identifying native snoRNA from pseudo ones, and we fold the candidates with our algorithm instead of RNAfold, which is used in snoReport (Hertel et al., 2008). Experiments have shown that our mining method outperforms snoReport (Zou et al., 2009a).

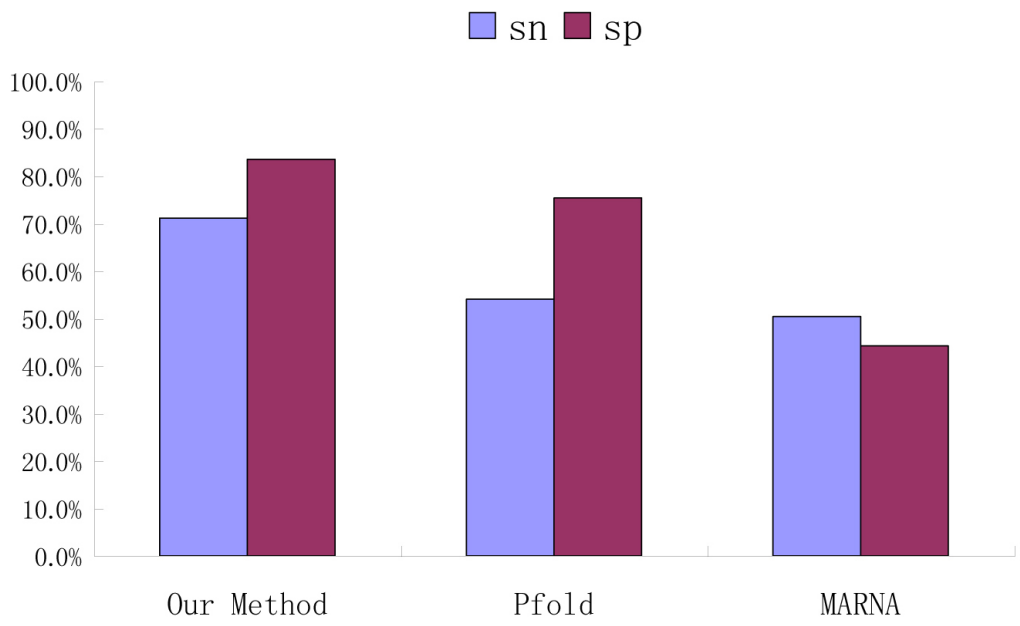

Figure 11. Performance comparison on multiple-sequence folding of RNAse P RNA. sn = TP / (TP + FN); sp = TP / (TP + FP).

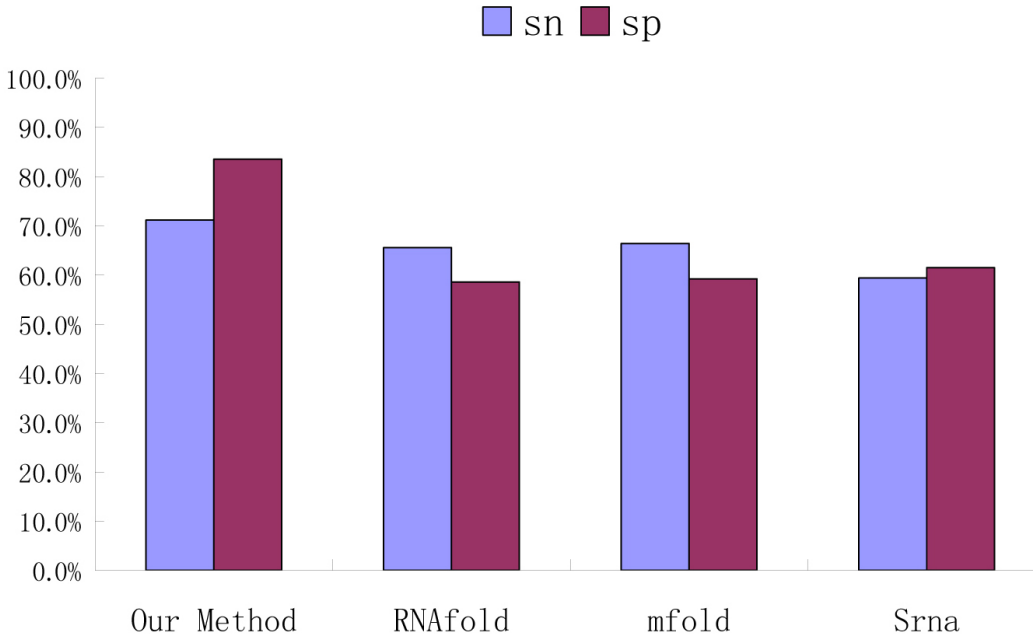

Figure 12. Performance comparison on single-sequence folding of RNAse P RNA. sn = TP / (TP + FN); sp = TP / (TP + FP).

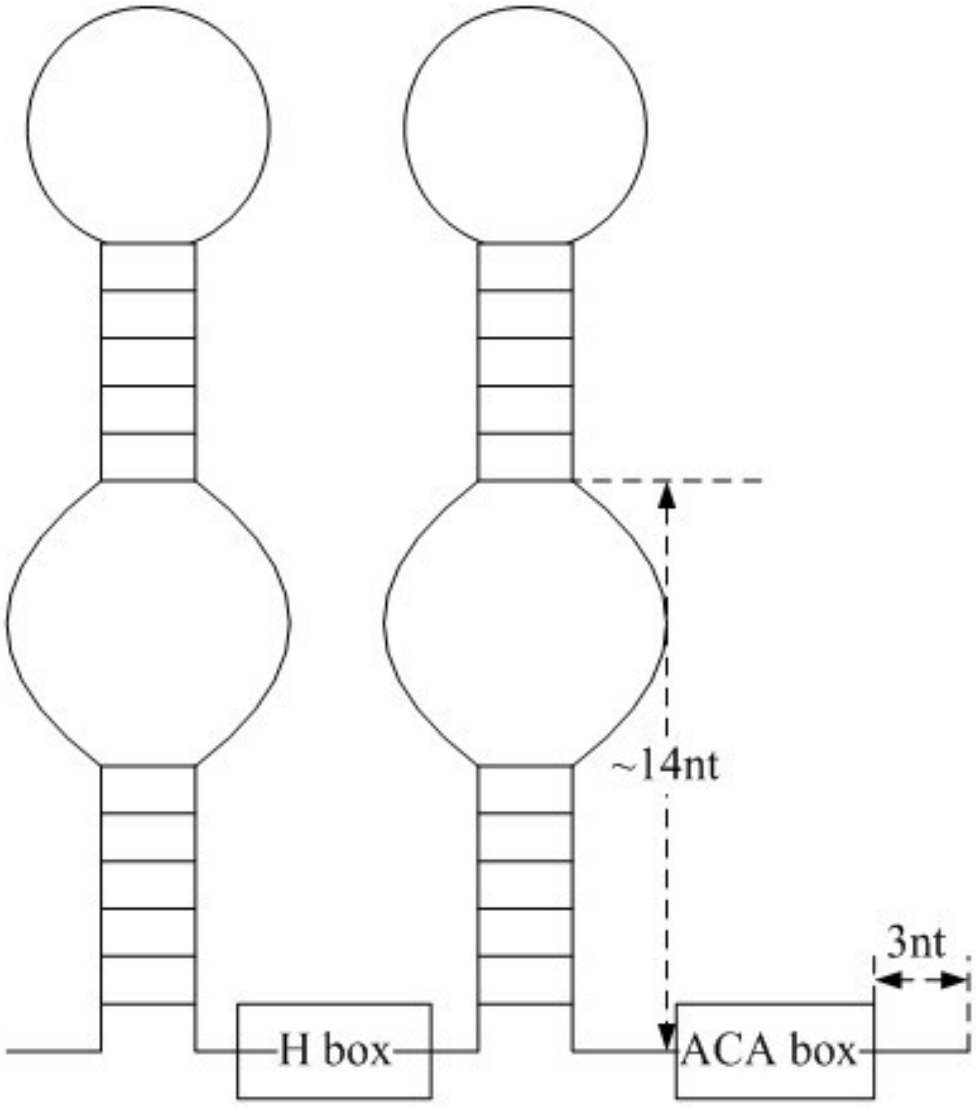

Figure 13. Schematic of the consensus structure of H/ACA box snoRNA.

The prediction method in this paper is also used in localizing the mature part of a microRNA precursor. This is a key problem in *ab initio* mining of microRNA. As is known, a microRNA precursor folds as a hairpin to keep low energy. Some precursors are proved by Northern blot analysis or RT-PCR; however, they are not folded as a hairpin by RNAfold or RNAStructure. This will influence further analysis. Our above algorithm replaces RNAfold when we perform studies on localizing the mature part and achieves a more accurate result (Zou et al., 2010).

In this paper, we propose a novel representation of RNA secondary structure and develop new folding algorithms. Experiments and application prove their performance. The

software is available at http://nclab.hit.edu.cn/~zouquan/zqfold/ or http://59.77.16.75/main/software/zqfold/index.htm.

ACKNOWLEDGMENTS

Research supported by the Natural Science Foundation of China (under grant #61001013, #60932008 and #61172098).

REFERENCES

Brown JW (1999). The ribonuclease P database. *Nucleic Acids Res.* 27: 314.

Byun Y and Han K (2006). PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Res.* 34: W416-W422.

Chan PP and Lowe TM (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* 37: D93-D97.

Ding Y, Chan CY and Lawrence CE (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.* 32: W135-W141.

Gardner PP and Giegerich R (2004). A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* 5: 140.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, et al. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33: D121-D124.

Hertel J, Hofacker IL and Stadler PF (2008). SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics* 24: 158-164.

Hofacker IL (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.* 31: 3429-3431.

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

Lambert A, Fontaine JF, Legendre M, Leclerc F, et al. (2004). The ERPIN server: an interface to profile-based RNA motif identification. *Nucleic Acids Res.* 32: W160-W165.

Larkin MA, Blackshields G, Brown NP, Chenna R, et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23: 2947-2948.

Leland W (1999). Dot Plots. *Am. Statistician* 53: 276-281.

Lowe TM and Eddy SR (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25: 955-964.

Rounsevell R, Forman JR and Clarke J (2004). Atomic force microscopy: mechanical unfolding of proteins. *Methods* 34: 100-111.

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

Touzet H and Perriquet O (2004). Carnac: folding families of related RNAs. *Nucleic Acids Res.* 32: W142-W145.

Witwer C, Hofacker IL and Stadler PF (2004). Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1: 66-77.

Zhang TT, Guo M and Zou Q (2007). RNA Secondary Structure Prediction Based on Forest Representation and Genetic Algorithm. Proceedings of the Third International Conference on Natural Computation, IEE Computer Society, Washington, 370-374.

Zou Q, Guo MZ, Liu Y and Xing ZA (2008). A Novel Comparative Sequence Analysis Method for ncRNA Secondary Structure Prediction Without Multiple Sequence Alignment. Proceedings of the Fourth International Conference on Natural Computation. IEE Computer Society, Washington, 29-33.

Zou Q, Guo MZ, Wang CY and Han YP (2009a). Novel H/ACA Box snoRNA Mining and Secondary Structure Prediction Algorithms. Proceedings of the Rough Sets and Knowledge Technology, Gold Coast, 538-546.

Zou Q, Zhao T, Liu Y and Guo M (2009b). Predicting RNA secondary structure based on the class information and hopfield network. *Comput. Biol. Med.* 39: 206-214.

Zou Q, Guo M, Liu Y and Xuan P (2010). DuplexFinder: predicting the miRNA-miRNA* duplex from the animal precursors. *Int. J. Bioinform. Res. Appl.* 6: 69-81.

Zuker M (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31: 3406-3415.

