

# MicroRNAfold: microRNA secondary structure prediction based on Modified NCM model with thermodynamics-based scoring strategy

Dianwei Han, Department of Computer Science, University of Kentucky  
Lexington, KY 40506-0046, dianweih@csr.uky.edu

Jun Zhang, Department of Computer Science, University of Kentucky  
Lexington, KY 40506-0046, jzhang@cs.uky.edu

and Guiliang Tang, Department of Plant and Soil Sciences,  
University of Kentucky  
Lexington, KY 40546-0236, gtang2@uky.edu

Technical Report CMIDA-HiPSCCS 010-08, Department of Computer Science,  
University of Kentucky, Lexington, KY, 2008.

## Abstract

MicroRNAs (miRNAs) are newly discovered endogenous small non-coding RNAs (21-25nt) that target their complementary gene transcripts for degradation or translational repression. In animals and plants, microRNAs play very important roles in cell growth, development and death. The biogenesis of a functional miRNA is largely dependent on the secondary structure of the miRNA precursor (pre-miRNA). An accurate prediction of the pre-miRNA secondary structure is important in miRNA informatics. For many years, thermodynamics-based methods have been the dominant strategy for single-stranded RNA secondary structure prediction. Recently, probabilistic-based methods have emerged to replace the free energy minimization methods for modeling RNA structures. However, the accuracies of the currently available best probabilistic-based models have yet to match those of the best thermodynamics-based methods. So this situation motivates us to develop a new prediction algorithm which will focus on microRNA structure prediction with high accuracy. A new model, nucleotide cyclic motifs (NCM), was recently proposed by Major *et al.* to predict RNA secondary structure. We propose and implement a novel model based on a Modified NCM (MNCM) model with a physics-based scoring strategy to tackle the problem of microRNA folding. Our MicroRNAfold is implemented by making use of a global optimal algorithm based on the bottom-up local optimal solutions. Our experimental results show that MicroRNAfold outperforms the current leading prediction tools in terms of True Negative rate, False Negative rate, Specificity, and Matthews coefficient ratio.

## Index Terms

Thermodynamics-based scoring function, RNA folding, MicroRNA secondary structure prediction, bottom-up local optimal solutions.

## I. INTRODUCTION

MicroRNAs (miRNAs) are newly discovered endogenous small non-coding RNAs (21-25nt) that are derived from larger hairpin RNA precursors and target their complementary gene transcripts for degradation or translational repression. MicroRNAs are found to play an important role in regulation of gene expression in plants and animals. Biologists assume that mammals have thousands of microRNAs in their genomes. MicroRNAs are expressed at different levels in animal and plant cells during cell differentiation, apoptosis, growth, and development. Understanding of microRNA pathways and microRNA biogenesis is considered to be a crucial aspect in tools development for functional genomics and metabolic engineering (Tang *et al.*) [20]. While the primary sequence information helps people understand miRNA pathways, in-depth understanding

of structure-function relationships requires knowledge of three-dimensional (3D) structure. It is very difficult and time-consuming to determine three-dimensional structures for natural RNA molecules. In addition, it has been reported that miRNA genes are more conserved in the secondary structure than in the primary sequences [1]. Secondary structural features should be more fully exploited in the homologue search for new miRNA genes. RNA secondary structure can be predicted with some accuracy by computers and many bioinformatics applications use certain notions of secondary structure in the analysis of RNA.

Generally there are two classes of algorithms available to predict the secondary structure of RNAs. The first class is the prediction methods based on phylogenetic sequence comparison, represented by Covariation prediction (Eddy *et al.*) [4] and Stochastic context free grammars (SCFG, Sakakibara *et al.*) [19], [3]. Both of them are based on a probabilistic model and an assumption that large numbers of homologous sequences from different organisms are available. When not enough related sequences are available, however, the second class of methods must be used, which are based on thermodynamics and represented by free energy minimization (Zuker *et al.*) [25], [24] and partition function method (McCaskill *et al.*) [16].

We examined several different theoretical strategies and studied their merits. Recent attempts to replace thermodynamics by statistical scores [2] led to similar or only slightly improved predictive power. More recently, Major *et al.* [17] proposed a new approach termed nucleotide cyclic motif (NCM), and developed MC-FOLD software to predict RNA structures. However, MC-FOLD can only deal with short RNA input sequences. In addition, there are several other RNA structural prediction software packages, such as Vienna package by Hofacker *et al.* [10], Mfold by Zuker *et al.* [24], CONTRAfold by Do *et al.* [3]. To better and more specifically predict miRNA secondary structure, the following aspects remain to be further investigated:

- (1) The currently available leading prediction tools are designed for general RNA structure prediction, which do not consider much the features of the microRNA secondary structures. There is a need to develop a more specific tool for miRNA secondary structure prediction.
- (2) While the currently available leading prediction tools achieve good accuracies on true positive cases, their accuracies on Matthews coefficient ratio [15] are relatively low.

Based on these observations, we decide to develop a new approach that will specifically deal with the microRNA secondary structure prediction.

We propose a new microRNA secondary structure prediction method based on Modified

NCMs (MNCMs), which makes use of thermodynamics-based scoring function, implemented as a computer program: MicroRNAfold. MicroRNAfold employs a bottom-up algorithm to compute many local optimal solutions. The global optimal solution is produced by sorting these local optimal solutions. Our experimental results show that our algorithm is very efficient in predicting MicroRNA secondary structure. In the future, we will build a 3D structure model based on the secondary structure prediction.

The organization of the paper is as follows: In Section II, we introduce our MicroRNAfold model, a global optimal algorithm based on bottom-up local optimal solutions, and some metrics used in our study. The experiments are carried out and the results are presented in Section III. A brief discussion is given in Section IV. We sum up this paper in Section V.

## II. METHODS

In this section, firstly, we demonstrate the use of MNCMs for RNA secondary structure prediction by showing how it arises as a natural extension of the recently developed NCMs. Secondly, we present MicroRNAfold, which is the hybrid model of traditional energy-based scoring schemes and MNCM structures. Finally, we introduce a global optimal algorithm which is based on the bottom-up local optimal solutions in our MicroRNAfold model.

### A. Modifying the definition of NCMs

In the work of Major *et al.*, NCM database contains lone pair loops up to six nucleotides [17]. For lone pair loops, they use the syntax “L- <sequence>”, where L is the length of the loop and <sequence> is the sequence. There are 4 types and 5440 lone pair loops: 64 3-loops (3-AAA, 3-AAC, ..., 3-UUU); 256 4-loops; 1024 5-loops; and 4096 6-loops.

Compared to the definition of standard NCMs [17], we make some modifications. We change the definition of lone pair based on the specific properties of microRNAs and the requirement of our algorithm. A valid lone pair structure must meet the following constraints:

(1) the first nucleotide and the last nucleotide in a lone pair must be Watson-Crick base pair or wobble base pair (G·U or U·G). We propose a new term *interface*, which is the boundary pair between two MNCMs. There is one interface for a lone pair MNCM. As shown in Fig. 1, G·C is the interface of a lone pair MNCM 5'GAACAC 3'. It is obvious that the first nucleotide and the last nucleotide in a lone pair form the interface. On the other hand, a double stranded

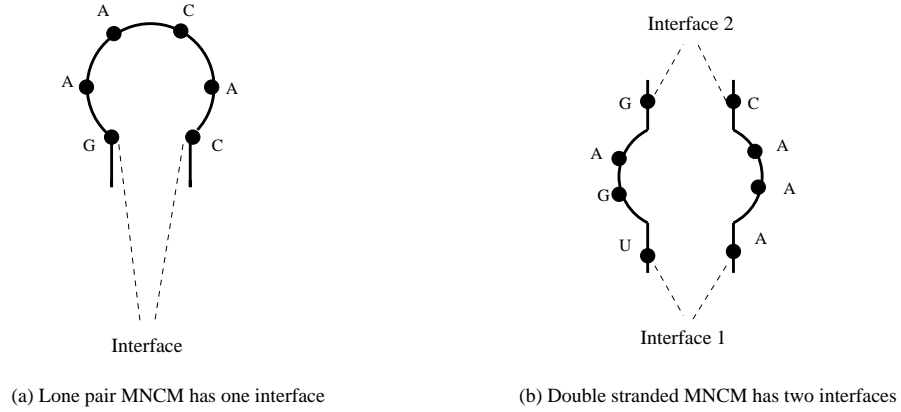


Fig. 1. **The definition of interfaces for MNCMs.** The pair G·C is the only interface for the lone pair MNCM (a). The pair U·A is an interface for (b) and the pair G·C is another interface for (b).

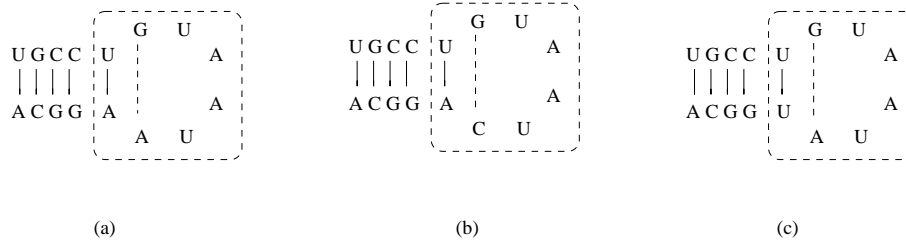


Fig. 2. **The selection of valid lone pair.** (a) is a valid lone pair. (b) is an invalid lone pair. (c) is an invalid lone pair.

MNCM has two interfaces. Consider a double stranded MNCM  $\begin{matrix} 5'UGAG3' \\ 3'AAAC5' \end{matrix}$  (see Fig. 1b) as an example, the pair U·A is interface1 and the pair G·C is interface2. In order to effectively employ our algorithm, we assume that all the interfaces should be a canonical base pair.

(2) the second unpaired pair is considered as the first mismatched pair of traditional minimal free energy algorithm. The second pair of a lone pair MNCM is the first mismatched pair of the hairpin loop according to the traditional thermodynamics-based models.

(3) the length of a lone pair ranges from 4 to a half of the length of the given sequence. This constraint is based on our experimental experience and pre-miRNA structure. In fact, the hairpin loop of a pre-miRNA may be very long and it contains far more than 6 nucleotides.

The definition of a lone pair is different from the one in MC-FOLD [17], and is not the same as the hairpin loop in traditional Minimal Free Energy (MFE) either. Fig. 2 depicts the selection of a valid lone pair. Let us focus on the parts (hairpin loop) within the dotted line rectangle. (a)

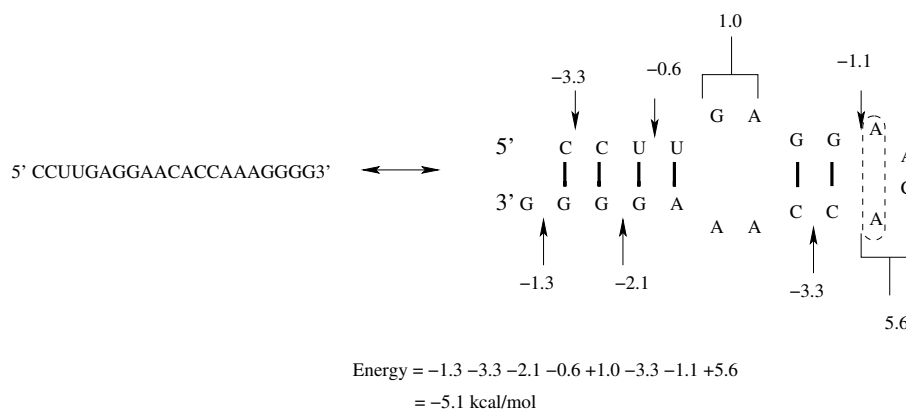


Fig. 3. **Prediction of conformational free energy for an RNA.** The total free energy is the sum of each increment.

is a valid lone pair according to our preset rules. (b) is an invalid lone pair because the second pair G·C behind the first pair U·A is a Watson-Crick base. (c) is an invalid lone pair because the first pair U·U is not a Watson-Crick base pair.

We use the same definition of double-stranded NCMs as Parisien *et al.* [17] do.

### B. From energy-based models to MNCMs

In order to describe the traditional energy-based model, we use an example.

Fig. 3 depicts the computation and model of free energy. This example is from Mathews *et al.* [13] and a similar strategy was adopted by Xia *et al.* [22]. The hairpin loop of four nucleotides AACA has an initiation of 5.6 kcal/mol. The dangling end (3'-most G) provides -1.3 kcal/mol of stability. The first mismatched pair A·A within dotted line in the hairpin loop is worth -1.1 kcal/mol. The 2x2 internal loop here destabilizes the structure, and its score is positive 1.0 kcal/mol.

The data structure that we use for our MicroRNAfold model is MNCMs. However, we use the experimentally measured thermodynamic parameters as scores instead of probabilities of each motifs or their operations [17]. We use the same example to present how to convert an energy-based model to an MNCM.

Fig. 4 describes the procedure of construction of items for the structure of an RNA.

(1) Construction of a lone pair (item m): 5'GAACAC 3'. As we mentioned earlier, the definition of our lone pair is different from that in the traditional energy-based model. The scoring function [22], [14] for a lone pair is:

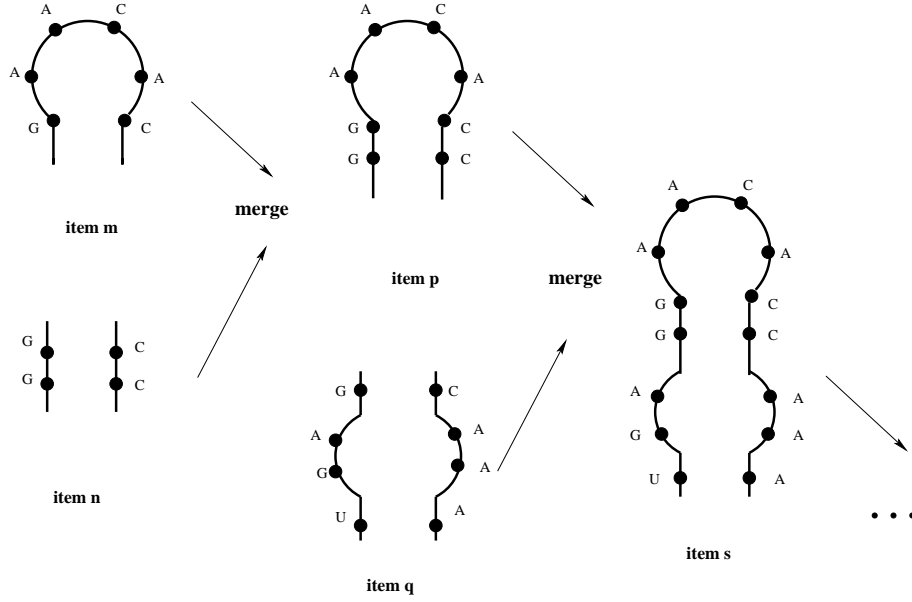


Fig. 4. The construction of items (cycles) in the MNCM model.

$$f(\text{ lone pair}) = \begin{cases} \Delta G_{37\text{initiation}(n)}^o + \Delta G_{37}^o(\text{ stacking of the first mismatch}) \\ + \Delta G_{37\text{bonus}}^o(\text{ U}\cdot\text{U or G}\cdot\text{A first mismatch, but not A}\cdot\text{G}) \\ + \Delta G_{37\text{bonus}}^o(\text{ special G}\cdot\text{U closure}) \\ + \Delta G_{37\text{penalty}}^o(\text{ oligo-C loops}) \end{cases}$$

(2) Construction of a double stranded MNCM (item n):  $\begin{matrix} 5' \text{GG} 3' \\ 3' \text{CC} 5' \end{matrix}$ . For this double helix, we use the identical scoring rule as Mathews *et al.* [22], [14] do.

(3) Merge item m and item n into item p: 5' GGAACACC 3'. We update the total score.

(4) Construction of a double stranded MNCM (item q):  $\begin{matrix} 5' \text{UGAG} 3' \\ 3' \text{AAAC} 5' \end{matrix}$ . We use the same scoring function as the one used by Xia *et al.* [22] and Mathews *et al.* [14] for this tandem mismatches.

(5) Merge item p and item q into item s: 5' UGAGGAACACCAAA 3'.

### C. MicroRNAfold modeling

The MicroRNAfold program applies MNCMs for RNA secondary structure prediction. The features in MicroRNAfold include:

- (1) base pairs,
- (2) helix closing base pairs,
- (3) hairpin lengths,
- (4) bulge loop lengths [5], [9], [11],
- (5) internal loop lengths,
- (6) internal loop asymmetry,
- (6) terminal mismatch interactions, and
- (7) dangling end.

Based on the features of the microRNA structures, we do not deal with pseudo knots and multi-branch loops.

1) *Generic base pairs*: In order to shorten our parameter tables and simplify our model, we merge canonical base pairs and terminal mismatches into one category: base pair. In fact we just consider mismatches as non-canonical base pairs.

2) *The 1x1 internal loops*: Due to the fact that we could not get all the needed parameters under this category, we just give some estimated values except for those publicly available data.

3) *The 2x2 internal loops*: We merge A·U/U·A cases and G·U/U·G cases together. We construct the table according to the publicly available data, and in other cases, just give the estimated value 2.8.

#### *D. Global optimal algorithm based on bottom-up local optimal solutions*

We implement our MicroRNAfold by a new recursive algorithm instead of Waterman-Byers algorithm [21]. We solve the problem by a backtracking method.

##### **The Algorithm:**

1. for all possible starting stems
2. Do:
  3. select one starting stem
  4. for all possible items for the next MNCM
  5. Do:
    6. Construct an item as the next MNCM
    7. if this item is satisfied with preset rules, the current item

- is added into the current stem, else try the next item
8.        endDo
  9.        construct a complete structure
  10.       repeat the above steps to get a local optimal solution
  11.       endDo
  12.       sort the sub-optimal structures
  13.       obtain a global optimal solution from many possible sub-optimal structures

In step 7, if the current item is satisfied with our preset rules which are different from the Waterman-Byers condition, we will add this current candidate item into the current stem. One new item is added each time. Meanwhile, the total score needs to be adjusted. A backtracking technique is used here, which is a bottom-up algorithm.

As shown in Fig. 5, the bottom-up (BU) algorithm is introduced by using an example. We start with a lone pair depicted in Fig. 5a. Then we repeatedly select a valid item as an MNCM and add this MNCM to the current structure until we construct a complete structure (see Fig. 5b1). At this moment, the stack pointer is at the beginning. We consider the beginning as the bottom and the lone pair as the top or head. We backtrack to the previous MNCM and rebuild the next possible structure (see Fig. 5b2). When we compare (b2) to (b1), we notice that the shadowed part is modified. When we go deep toward the lone pair, we can construct the structures shown in Fig. 5bi and in Fig. 5bn. The local optimal structure with the minimum score among the candidates (b1, b2, ..., bi, ..., bn) is chosen. Based on the different lone pairs, we obtain many different local optimal structures. The global optimal solution is obtained by applying insertion-sort algorithm.

### *E. Accuracy metrics*

In order to precisely assess the predictive power of prediction methods, we use some typical measures, which have been extensively applied in the field of bioinformatics. Measures used in our study include True Positive rate, False Positive rate, True Negative rate, and False Negative rate, in addition to some more important metrics *Matthews*, *Sensitivity*, and *Specificity*.

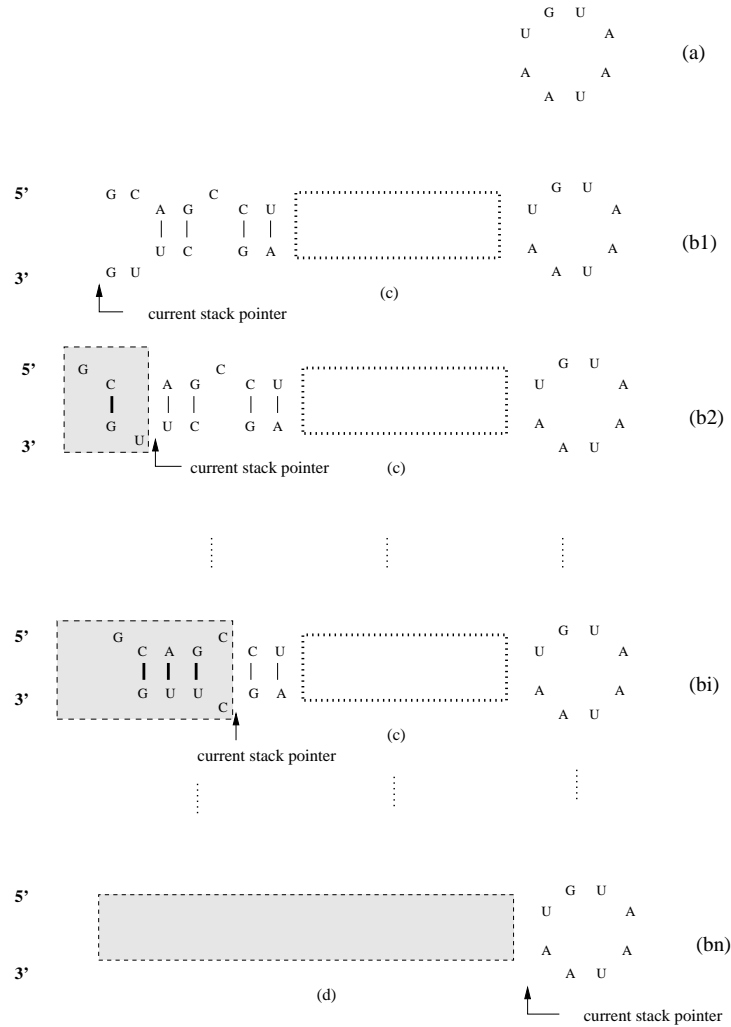


Fig. 5. **The bottom-up algorithm.** (a) denotes a lone pair (the hairpin loop). (b1) displays the first initial structure that the program constructs with a given input sequence. (b2) denotes another structure when we backtrack the stack pointer. (bi) denotes that at the  $i$ th step, this structure is produced by the program. (bn) denotes the last structure that is built by the program. (c) shows the part of structure that remains unchanged from (b1) to (bi). (d) is the stem part of the last structure based on current lone pair. Compared to the previous structures, the modified part is shown by the shadowed area.

*Sensitivity* can be defined as

$$Sensitivity = \frac{\text{number of correct base pairings}}{\text{number of true base pairings}}. \quad (1)$$

We see that sensitivity is equal to True Positive rate here.

*Specificity* can be defined as

$$Specificity = \frac{\text{number of correct base pairings}}{\text{number of predicted base pairings}}. \quad (2)$$

> MicroRNAfold (input file is from dps-mir-284 sequence)

Top 10 scores:

```

----- RESULT -----
Sequence : GUUGCAGUCCUGGAAUUAAAGUUGACUGUGUAGCCUGGGAAGGCAAGGCUUGAGCACUGCUUCUGAAGUCAGCAACUUGAUCCAGCAAUUGCGGCCAA
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -15.473984
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -14.343483
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -10.783984
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -10.143984
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -7.723985
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -6.823986
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -5.643985
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -5.187428
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -5.063985
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))))... -3.687425

```

Fig. 6. **MicroRNAfold predictions for dps-mir-284.** The top ten structures generated by MicroRNAfold for dps-mir-284. The structures are shown in dot-bracket notation. A parenthesis represents a canonical base pair; a dot represents an unpaired nucleotide. A dot-bracket can be converted in a secondary structure representation. Negative floating point numbers on the right hand side denote the corresponding scores.

Matthews coefficient ratio [15] can be written as:

$$Matthews = \sqrt{\frac{TP \times TP}{(TP + FN) \times (TP + FP)}} \quad , \quad (3)$$

where FP is the number of false positive cases, FN is the number of false negative cases, and TP is the number of true positive cases.

### III. RESULTS

We evaluated the predictive power of MicroRNAfold by using known secondary structures of non-coding RNA taken from the miRBase database [6], [7], [8]. Our testing data set comes from *Arabidopsis thaliana*, *Brassica napus*, *Saccharum officinarum*, *Homo sapiens*, *Gallus gallus*, *Glycine max*, *Apis mellifera*, *Drosophila melanogaster*, and *Drosophilla pseudoobscura*. The sequence lengths of the testing data set range from 59 to 188. We implemented the MicroRNAfold by using ANSI C code and the program was ran on a Linux-based machine. We used Pseudoviewer to view our structures [18]. Our best solution is from the first one among several hundreds of sorted possible structures (see Fig. 6).

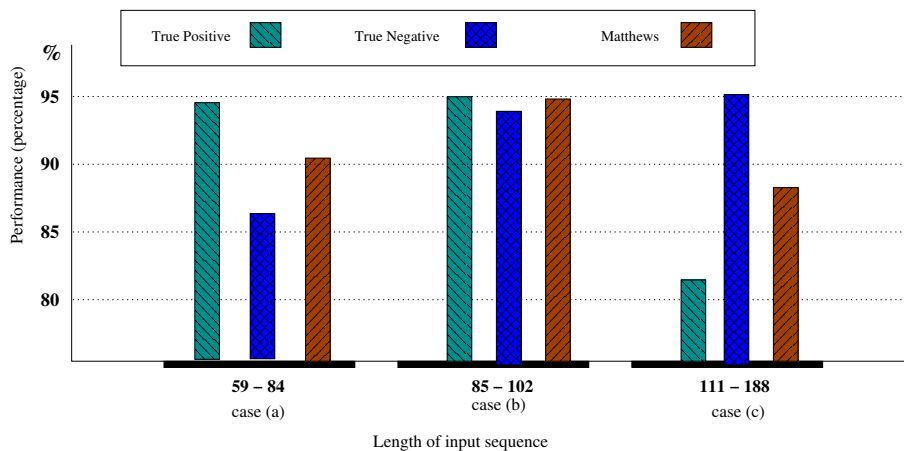


Fig. 7. **The specific performance based on different sequence lengths.** We divide the sequence lengths into three groups: case(a) (59-84), case(b) (85-102), and case(c) (111-188).

#### A. Predictive power of *MicroRNAfold* associated with different sequence lengths

In our study, we considered only AU, GC, and GU base pairs because there is no sufficient knowledge concerning the non-canonical base pairs even though non-canonical base pairs might be important and play some roles in determining 3D structures of RNAs [17].

Fig. 7 shows the specific performance based on different sequence lengths. It shows that our *MicroRNAfold* does best when the lengths of the input sequences range from 85 to 102 which are the average lengths for normal microRNAs. The Matthews value in case (b) increases to 94.50% from 90.84%, compared to case (a). Compared to case (a), the Matthews value in case (c) drops a little bit but the True Positive rate drops significantly. It makes sense because it is always a challenge for a prediction approach when the input sequence is very long. On the other hand, the True Negative rate in case (c) increases a little, compared to case (a) and case (b).

#### B. Predictive power of *MicroRNAfold* associated with different hairpin loop lengths

Fig. 8 depicts the comparison of the *MicroRNAfold* performance based on the hairpin loop lengths. We just use the Matthews value as the metric to assess the effectiveness and the performance of our system. It seems that *MicroRNAfold* obtains the best Matthews coefficient ratio when the hairpin loop is in average length. This result implies that maybe there is a relationship between the accuracies of the prediction approaches and the lengths of the hairpin loop of miRNAs.

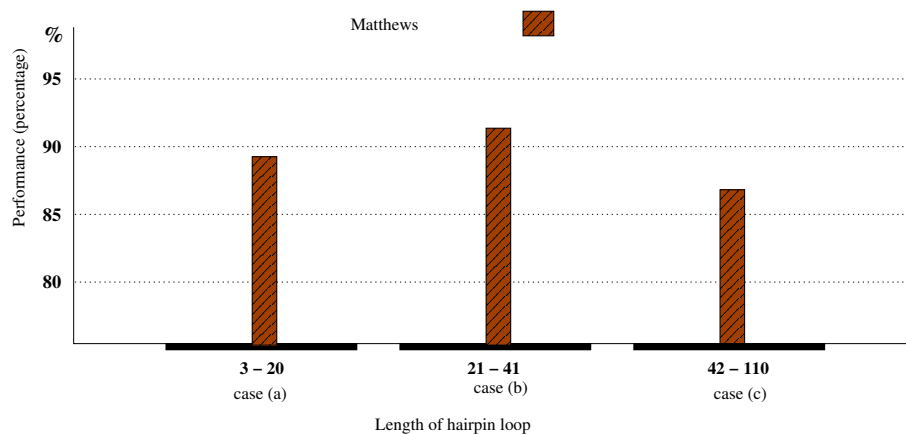


Fig. 8. **The Matthews coefficient ratio performance based on different hairpin lengths.** We divide the hairpin loop lengths into three groups: case(a), case(b), and case(c). The case(a) indicates that the length range is 3-20, the case(b) indicates that the length range is 21-41, and the case(c) indicates that the length range is 42-110.

To see more examples, please read Appendix I.

### C. Comparison to other methods

We compared the performance of MicroRNAfold with the other two leading methods: one adopts the probabilistic-based strategy, and the other chooses the free energy minimization strategy. For benchmarking experiments, we used MC-FOLD [17], and Mfold (<http://bioinfo.hku.hk/Pise/mfold.html>) [24], with default parameters for each program. All benchmarks were conducted on Intel-based servers running a GNU/Linux operating system. Whenever a program returned multiple possible structures (e.g., Mfold and MC-FOLD), we chose the structure with the minimum score.

Fig. 9 shows the comparison of the predictive power of different methods. Compared to the thermodynamic approach and the probabilistic method with NCM, MicroRNAfold obtains a higher Matthews coefficient ratio, a higher True Negative rate and a lower False Negative rate, despite a lower True Positive rate and a higher False Positive rate. In particular, MicroRNAfold achieves statistically significant improvements of over 12% in specificity relative to the best current method, Mfold. In the aspect of the True Positive rate, MC-FOLD and MFold work better than MicroRNAfold.

Prediction Methods:	MC-FOLD (NCM)	MFold (Thermodynamics)	MicroRNAfold (Modified NCM plus thermodynamics)
<b>True Positive rate</b> = Sensitivity	95.00 %	<b>98.91 %</b>	90.37 %
<b>False Positive rate</b>	5.00 %	<b>1.09 %</b>	9.63 %
<b>True Negative rate</b>	26.91 %	65.91 %	<b>93.02 %</b>
<b>False Negative rate</b>	72.82 %	34.09 %	<b>6.98 %</b>
<b>Specificity</b>	75.41 %	81.57 %	<b>94.49 %</b>
<b>Matthews coefficient ratio</b>	73.34 %	85.76 %	<b>91.59 %</b>

Fig. 9. **Comparison of the predictive power with other prediction methods.** The predictions are compared over 1101 base pairs. For each approach, the best predicted structures are analyzed. In each row, we use bold font to represent the best value.

#### IV. DISCUSSION

Although we have obtained encouraging results compared to other prediction approaches, there are still some issues that need to be discussed in detail. The first thing that we would like to mention is the auxiliary information. As we know, all the parameters and understanding of RNA secondary structure come from experimental results. Experimental results and the related analysis based on the experimental facts may help us design a more accurate model and prediction algorithm. How to get this knowledge is still a challenge for us. The second thing is the scoring strategy. During our testing phase, we found some proposed structures from the database could not be generated from our results based on the current scoring function.

##### A. Taking into account auxiliary information and more parameters

Sometimes we could not successfully predict the secondary structure of an RNA because “our knowledge of the contributions of various RNA motifs to the total free energy of RNA structures is still incomplete” [22]. Due to the limitation of this kind of knowledge, we could not give all the thermodynamic parameters concerning free energy. Thus it could affect our prediction negatively. For example, when we predicted the microRNA hsa-mir-196a, we failed to achieve

the proposed structure of 
$$\begin{array}{l} 5'UUAG3' \\ \quad \cdot \\ 3'AGCC5' \end{array}$$
.

Fig. 10 shows the comparison of the predicted structure by MicroRNAfold with the structure proposed by the database. According to our current scoring function, the sum of the two parts

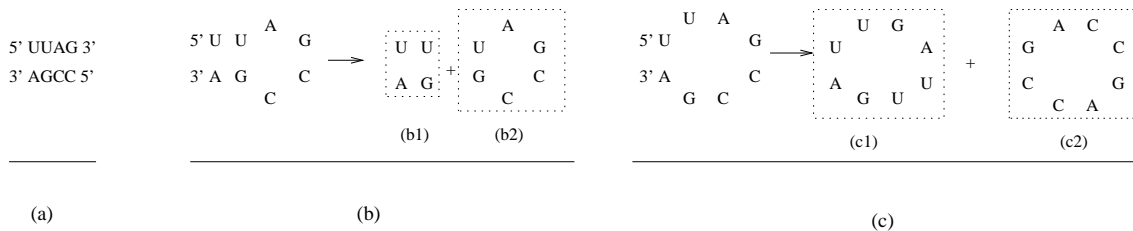


Fig. 10. **Prediction of a specific structure.** (a) is the sequence of the structure, (b) is the predicted structure by MicroRNAfold, and (c) is the proposed structure by the database.

(b1) and (b2) is  $-0.07$  and the score of the structure (c) is  $0.95$ . Therefore, we take (b) as the solution. When we calculate the score for the structure (c), we use the following formula proposed by Xia *et al.* [23]:

$$\Delta G_{37predict}^o(c) = [\Delta G_{37loop}^o(c1) + \Delta G_{37loop}^o(c2)] * \frac{1}{2} + \Delta \quad (4)$$

In order to solve this problem, we need to refine scoring function or incorporate auxiliary information. Based on the statistical and theoretical analysis of the experimental data, we may incorporate biological constraints to help the prediction [25].

### B. Some issues with scoring strategy

During our study, we found that some of the best structures did not come from the first structure whose score is minimal. For example, according to the miRBase database, the structure of ame-mir-317 should be (b) in Fig. 11. But the results from MicroRNAfold and Mfold both showed that the proposed structure should be (a) in Fig. 11.

Fig. 12 displays the different hairpins between the database and our prediction approach. We can see that the hairpin which we obtained by our method is the same as the one obtained by Mfold. Is that saying that our prediction is correct and the one in the database is wrong? Let us see our scores. According to our strategy, the score for the hairpin of (b) (in Fig. 12) is  $5.0$  while the score for the hairpin of (a) (in Fig. 12) is  $5.15$ . So, we choose (b) as the structure of the hairpin based on our current scoring function. In order to improve our prediction algorithm we have to refine the scoring function.



can consider building a 3D structure model based on the secondary structure prediction.

## APPENDIX I

### AN EXAMPLE OF PREDICTION CONCERNING DIFFERENT HAIRPIN LOOP LENGTHS

As shown in Fig. 13, we see an example. In case (a), the MicroRNAfold did not successfully predict the hairpin loop. In addition, MicroRNAfold did not identify the 2x2 internal loop 5'UG3' . In case (b), the MicroRNAfold successfully constructed the hairpin. As for the stem, 3'GU5'

MicroRNAfold did not predict the 2x2 internal loop  $\begin{matrix} 5'UA3' \\ 3'UG5' \end{matrix}$  and two canonical base pairs G·C and G·U. Compared to (a) and (b), case (c) is a little worse. The MicroRNAfold did not predict the 2x2 internal loop  $\begin{matrix} 5'UC3' \\ 3'UA5' \end{matrix}$  . In addition, It did not obtain the correct structure for the nucleotides 1-6 and the nucleotides 129-135.

## ACKNOWLEDGMENT

The authors would like to thank Dr. Chingwen Li for discussion about MicroRNAfold and for suggestions to improve the manuscript. This work was supported in part by USDA-NRI grants 2006-35301-17115 and 2006-35100-17433, NSF MCB-0718029, and in part by the Kentucky Science and Technology Corporation under Contract KSTC-144-401-08-029.

## REFERENCES

- [1] Pierre Baldi, Søren Brunak. Neural Networks: Applications. *Bioinformatics: The Machine Learning Approach*, MIT Press, Boston, MA, page 113-155 (2001)
- [2] Stephan H Bernhart, Hakim Tafer, Ulrike Mückstein, Christoph Flamm, Peter F Stadler, Ivo L Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for Molecular Biology*, 1:3(2006)
- [3] Chuong B. Do, Daniel A. Woods, Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90-e98 (2006)
- [4] Sean R. Eddy, Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079-2088 (1994)
- [5] Thomas R. Fink, Donald M. Crothers. Free energy of imperfect nucleic acid helices. I. The bulge defect. *J. Mol. Biol.*, 66:1-12 (1972)
- [6] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, Anton J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36: D154-D158 (2008)

- [7] Sam Griffiths-Jones, Russell J. Grocock, Stijn van Dongen, Alex Bateman, Anton J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34: D140-D144 (2006)
- [8] Sam Griffiths-Jones. The microRNA registry. *Nucleic Acids Research*, 32: D109-D111 (2004)
- [9] Duncan R. Groebe, Olke C. Uhlenbeck. Thermal stability of RNA hairpins containing a four-membered loop and a bulge nucleotide. *Biochemistry*, 28:742-747 (1989)
- [10] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, L. Sebastian Bonhoeffer, Manfred Tacker, Peter Schuster. Fast folding and comparison of RNA secondary structures (the Vienna RNA package). *Monatsh Chem.*, 125:167-188 (1994)
- [11] Carl E. Longfellow, Ryszard Kierzek, Douglas H. Turner. Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry*, 29:278-285 (1990)
- [12] David H. Mathews, D. H. Turner. Prediction of RNA secondary structure by free energy minimization. *Current Opinion in Structural Biology*, 16:270-278 (2006)
- [13] David H. Mathews, Michael Zuker. Predictive Methods Using RNA Sequences. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd ed.*, edited by Andreas D. Baxeavanis and B.F. Francis Ouellette, John Wiley & Sons, page 144-167 (2005)
- [14] David H. Mathews, Jeffrey Sabina, Michael Zuker. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911-940 (1999)
- [15] Brian W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Biophysica Acta*, 405:442-451 (1975)
- [16] John S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105-1119 (1990)
- [17] Marc Parisien, Francois Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452:51-55 (2008)
- [18] <http://pseudoviewer.inha.ac.kr/>
- [19] Yasubumi Sakakibara, Michael Brown, Richard Hughey et al. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22(23):5112-5120 (1994)
- [20] Guiliang Tang, Gad Galili, Xun Zhuang. RNAi and microRNA: Breakthrough technologies for the improvement of plant nutritional value and metabolic engineering. *Metabolomics*, 3:357-369 (2007)
- [21] Michael S. Waterman, Thomas H. Byers. A dynamic programming algorithm to find all solutions in the neighborhood of the optimum. *Mathematical Biosciences*, 77:179-188 (1985)
- [22] Tianbing Xia, David H. Mathews, Douglas H. Turner. Thermodynamics of RNA secondary structure formation. *In Prebiotic Chemistry, Molecular Fossils, Nucleotides, and RNA*, edited by D. G. So"ll, S. Nishimura, and P. B. Moore, Elsevier, page 21-48 (1999)
- [23] Tianbing Xia, J.A. McDowell, Douglas H. Turner. Thermodynamics of nonsymmetric tandem mismatches adjacent to G-C base pairs in RNA. *Biochemistry*, 36:12486-12487 (1997)
- [24] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406-3415 (2003)
- [25] Michael Zuker, Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133-148 (1981)



**Dianwei Han** received an M.S. degree from Lamar University, USA, in 2003. He is currently a Ph.D. student and a member of the Laboratory for Computational Medical Imaging & Data Analysis in the Department of Computer Science at the University of Kentucky, USA. His research interests include data mining and knowledge discovery, database analysis, and bioinformatics.



**Jun Zhang** received a Ph.D. from The George Washington University in 1997. He is a Full Professor of Computer Science and Director of the Laboratory for High Performance Scientific Computing & Computer Simulation and Laboratory for Computational Medical Imaging & Data Analysis at the University of Kentucky. His research interests include computational neuroinformatics, data mining and information retrieval, large scale parallel and scientific computing, numerical simulation, iterative and preconditioning techniques for large scale matrix computation. Dr. Zhang is associate editor and on the editorial boards of eight international journals in computer and computational sciences, and is on the program committees of a few international conferences. His research work has been funded by the U.S. National Science Foundation, the Department of Energy, and National Institutes of Health. He is recipient of the U.S. National Science Foundation CAREER Award and several other awards.



**Guiliang Tang** received his Ph.D. in 2001 at the Weizmann Institute of Sciences in Israel. He was a post-doctoral Fellow of the Charles A. King Trust Research Fellowship with the Medical Foundation at Boston. There, he worked on RNA Interference (RNAi) and microRNA (miRNA) with Dr. Phil Zamore of the University of Massachusetts Medical School. In July 2005, Dr. Tang started his Gene Suppression Laboratory as a faculty member and a Principal Investigator at the University of Kentucky. His research focuses on dissecting the mechanism of RNAi and miRNA, as well as developing tools of gene suppression for gene function and functional genomics. Dr. Tang also teaches a laboratory course on RNAi and miRNA for undergraduate and graduate students.

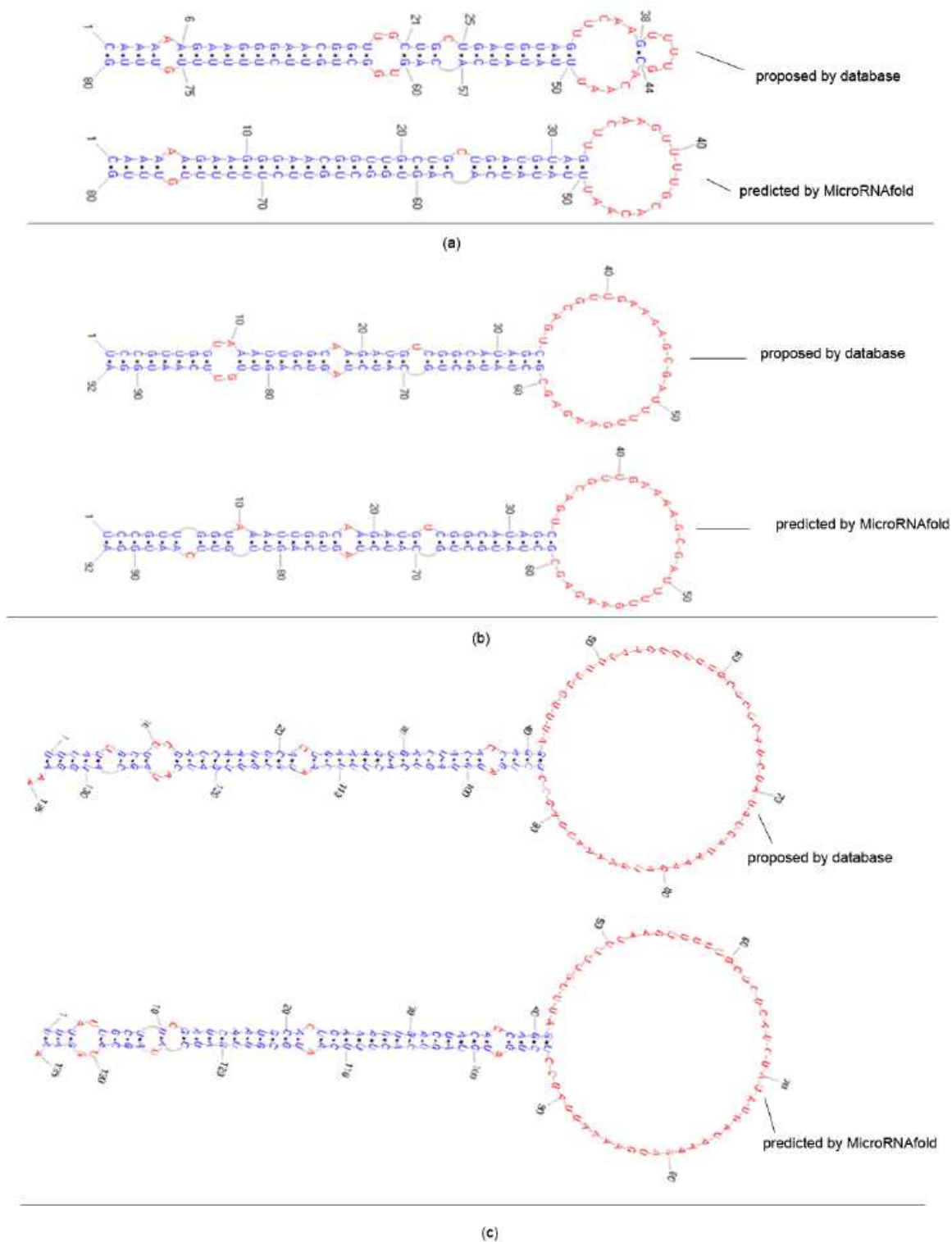


Fig. 13. **An example of prediction based on different hairpin lengths.** (a) depicts the structure proposed by database and the structure predicted by MicroRNAfold with the input of Mirna dsp-mir-6-3 sequence. In this case, the length of hairpin loop is 5. (b) shows the structure proposed by the database and the structure predicted by MicroRNAfold with the input of Mirna dme-mir-31a sequence. In this case, the length of hairpin loop is 27. (c) depicts the structure proposed by database and the structure predicted by MicroRNAfold with the input of Mirna bna-mir-161 sequence. In this case, the length of the hairpin loop is 52.