A New Method to Predict RNA Secondary Structure Based on RNA Folding Simulation

Yuanning Liu, Qi Zhao, Hao Zhang, Rui Xu, Yang Li, and Liyan Wei

Abstract—RNA plays an important role in various biological processes; hence, it is essential when determining the functions of RNA to research its secondary structures. So far, the accuracy of RNA secondary structure prediction remains an area in need of improvement. This paper presents a novel method for predicting RNA secondary structure based on an RNA folding simulation model. This model assumes that the process of RNA folding from the random coil state to full structure is staged and in every stage of folding, the final state of an RNA is determined by the optimal combination of helical regions, which are urgently essential to dynamics of RNA formation. This paper proposes the First Large Free Energy Difference (FLED) in order to find the helical regions most urgently needed for optimal final state formation among all the possible helical regions. Tests on the datasets with known structures from public databases demonstrate that our method can outperform other current RNA secondary structure prediction methods in terms of prediction accuracy.

Index Terms—Free energy, ribonucleic acid, RNA folding, RNA secondary structure prediction

1 Introduction

RNA molecules are an important component of biological substance; they are not only the carriers of genetic information between DNA molecules and proteins [1], but they also play an important role in many biological processes, such as catalysis [2], protein synthesis [3], [4], [5], immunity [6], development [7] and many other important biological processes. RNA molecules have a three-tier hierarchy structure, beginning with the primary sequence, then the secondary structure (Fig. 1), i.e., the set of base pairs, and ultimately the tertiary structure, i.e., the full three-dimensional structure. The functions of RNA molecules mainly rely on their tertiary structures. Because the time is much shorter and the free energy change is much larger during RNA secondary structure formation than during its tertiary structure formation, the secondary structure can be predicted independently of the tertiary structure [8]. Furthermore, secondary structure prediction is an important intermediate step needed to predict RNA tertiary structure [9]. Therefore, obtaining accurate RNA secondary structure is crucial to determining the functions of RNAs and the biological processes they are involved in.

At present, the experimental techniques for RNA structure determination, such as X-ray crystallography and Nuclear Magnetic Resonance (NMR), remain difficult, costly and time consuming. So capturing images of RNA secondary structure with the help of biological computing is still the preferred method.

Comparative sequence analysis is the most accurate method [10], and the positive predictive value (PPV) of this method exceeds 97 percent [11] when a group of homologous sequences are available. However, this method is not only time-consuming but also requires a plurality of homologous sequences and intensive human monitoring [12].

- *Y. Liu, Q. Zhao, H. Zhang, R. Xu, and Y. Li are with the Computer Science and Technology Department, Jilin University, Changchun 130000, Jilin, China. E-mail: {liuyn, zhangh}@jlu.edu.cn, qizhao13@mails.jlu.edu.cn, {419162090, 893142912 }@qq.com.*
- *L. Wei is with the Software Engineering Department, Jilin University, Changchun 130000, Jilin, China. E-mail: 308555989@qq.com.*

Manuscript received 7 Dec. 2014; revised 28 Mar. 2015; accepted 5 Aug. 2015. Date of publication 3 Nov. 2015; date of current version 4 Oct. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2496347

A free energy minimization method based on dynamic programming is the most common method used to predict the RNA secondary structure of a single sequence [12], [13], such as mfold [14]. This method assumes that under certain conditions, an RNA molecule adjusts itself to the structure with the minimum free energy (MFE), which is thermodynamically the most stable structure [15]. Therefore, the MFE structure is the native structure of an RNA. A dynamic programming algorithm can guarantee locating the global MFE structure, and it requires $O(N^3)$ time. The result of this method is better for predicting short sequences, but for long sequences such as full-length small subunit rRNA (ribosomal RNA) and large subunit rRNA, the sensitivity of this method is not satisfactory [16]. In addition, biological experiments have shown that the native RNA structures are not usually the MFE structures [17], and also the calculation method for free energy is not perfect [18].

Only the thermodynamic factor is taken into account in the free energy minimization method and MFE structure is assumed to be the native structure, while the process of RNA structure formation is overlooked. Studies have shown that differences between the prediction results of long RNAs secondary structures and their native structures cannot be simply classified as an error in thermodynamic parameters, which led to the conclusion that RNA folding is related to formation dynamics [19]. Proctor and Meyer improved the free energy minimization method by taking into account the kinetic factors during RNA folding [20]; moreover the sensitivity and PPV of the improved method has proven superior to RNAfold [15].

Methods based on a dynamic programming algorithm take the base pair as the basic unit of RNA secondary structure, but the formation of a helical region is the rate limiting step. In other words, once a few base pairs of a helical region are formed, the rest of the base pairs will be formed quickly [21]. Therefore, taking the helical region as the basic unit of RNA secondary structure is more theoretically feasible. Helical region distribution is used in these kinds of methods [22], [23], [24] to search for all possible helical regions from an RNA sequence, which compose the candidate helical region set. Any compatible non-empty subset of the candidate helical region set can be a secondary structure. In fact, the empty set refers to the RNA primary structure. The problem of an RNA secondary structure is then converted into the problem of how to select a subset from candidate helical region set. Among the possible solutions, [22] is the most typical method used for subset selection. This algorithm takes RNA folding as an iterative process, but it is also a greedy algorithm designed to simply choose the helical region that can reduce the most free energy of the structure in each iteration. However, the result of this algorithm is not satisfactory. The reason may be that it simply tries to find the thermodynamically minimum free energy, so that its result is often trapped in local minima.

Based on the problems of RNA secondary structure prediction, this paper proposes a new method based on the folding simulation model, which combines both thermodynamics and kinetics factors of the RNA secondary structure. Another novel approach adopted in this method is to not assume that the native structure of an RNA is always the MFE structure. This approach recongnizes that a native structure must be thermodynamically stable [17], and that the need for stability is closely related to the formation process. Hence, the RNA folding process from the random coil state to full structure is staged and in every stage of folding, the final state of an RNA is determined by the optimal combination of helical regions which urgently need to form under the current RNA state. This paper presents the First Large Free Energy Difference (FLED) in order to find these helical regions in urgent need of formation among all the possible helical regions. Tests on the datasets with known structures from public databases demonstrate that our method can outperform other current RNA secondary structure prediction methods in terms of prediction accuracy.

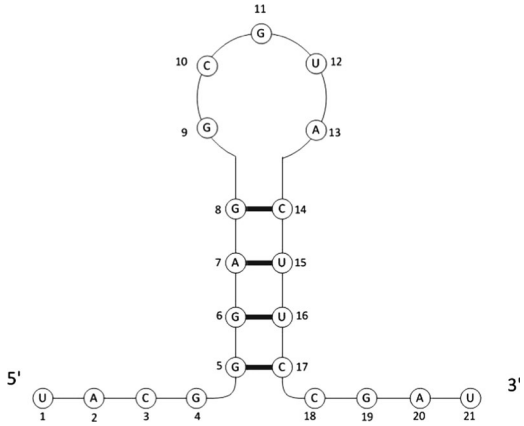

Fig. 1. RNA secondary structure diagram. Each circle represents a base whose number is indicated beside it. The length of this RNA is 21. Thick lines connect paired bases. (5, 8), (14, 17) represent two regions, and they can be paired reversely, so that (5,17,4) is a helical region whose helical region interval is 5.

2 Methods

2.1 Some Definitions

It is necessary to establish some definitions before proceeding.

An RNA sequence is composed of L $(L \geq 1)$ bases; then, the length of the RNA is L. Number the bases of the RNA consecutively from 1 to L beginning at the 5' end of the RNA and it can be expressed as:

$$sequence = b_1 b_2 \dots b_L, b_i \in \{A, C, G, U\}, 1 \leq i \leq L.$$

(n, m) is called a region indicating the nucleotide sequence from b_n to b_m, where $1 \leq n < m \leq L$. Actually, region $(1, L)$ represents the primary sequence of an RNA. For region (i, j) and region(p, q), they do not intersect if $j > p$ or $q > i$.

$b_i \bullet b_j$ denotes b_i paired with b_j, where $b_i \bullet b_j \in \{A \bullet U, U \bullet A, G \bullet C, C \bullet G, G \bullet U, U \bullet G\}, 1 \leq i < j \leq L$ $b_i \bullet b_j$ and $b_p \bullet b_q$ are compatible if $i \neq p, i \neq q, j \neq p, j \neq q$.

If all the bases of two regions which are of equal length and do not intersect each other can be paired reversely, the set of such base pairs is called a helical region, namely, for the region (i, j) and region (p, q) (assuming $j > p$), if $j - i = p - q, b_{i+a} \bullet b_{q-a}$, $a = 0, 1, \dots, j - i$, then the set $\{b_{i+a} \bullet b_{q-a}\}, a = 0, 1, \dots, j - i$ is a helical region. And it can also be expressed as a triplet: $h = (i, q, length)$; hence, b_i is the base closest to the 5' end, b_q is the base closest to the 3' end and $length = j - i + 1$ is the number of base pairs, namely the length of the helical region. The number of bases between two regions of a helical region is called helical region interval with $interval = i - q - 2length + 1$. Apparently, the helical region interval of a hairpin structure is equal to the number of bases on its hairpin ring. If $\forall b_i \bullet b_j \in h_1$ is compatible with $\forall b_p \bullet b_q \in h_2$, the helical region h_1 and h_2 are compatible.

The set composed of all possible helical regions in terms of an RNA sequence is a candidate helical region set H. So the RNA secondary structure can be defined as a non-empty compatible subset of H, namely, $Structure \subseteq H$, $\forall h_i$ and $\forall h_j$ are compatible, $h_i, h_j \in Structure, i \neq j, H \neq \emptyset$.

2.2 Rules for Searching All the Maximum Helical Regions

For a helical region h, usually the contact matrix $C_{L \times L}$ can be constructed by helical region distribution, where L is the length of h. Contact matrix $C_{L \times L}$ is initialized to be an empty matrix and $C_{i,j}$ is marked if b_i is paired with b_j. Then by searching the matrix $C_{L \times L}$, all the possible helical regions can be obtained and they compose the original candidate helical region set H. We then make further restrictions on the original candidate helical region set H.

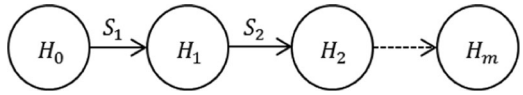

Fig. 2. State transition in the staged folding process.

$\forall h_i \in H$, $h_i = (start, end, length)$, $interval_i$ is the helical region interval of h_i and h_i should comply with:

1. $length \geq 3$;
2. Both base pairs at the end of h_i, $b_{start} \bullet b_{end} \neq G \bullet U$, $b_{start+length-1} \bullet b_{end-length+1} \neq G \bullet U$;
3. $interval_i \geq 3$;
4. No other helical region, which complies with rules 1-3, contains all the base pairs in h_i.

All the rules above are derived from the native RNA secondary structures and some appropriate simplifications are made. In fact, these rules are designed to pick out the helical regions which are most likely to exist in their native structures, and the number of them should be as small as possible to speed up the execution of the program. The fourth rule is to pick up the maximum helical regions in all the possible helical regions. If the length of the maximum helical region is L, the helical region may contain $(L^2 - 3L)/2$ other helical regions. Obviously, selecting only the maximum helical regions greatly reduces the number of helical regions.

After screening as per the above limitations, the final candidate helical region set can be obtained. Of course, in the native structures, some helical regions do not exist as the maximum helical regions.

2.3 The RNA Folding Simulation Model

For the RNA secondary structure, base stacking between base pairs is one of the main factors that make the RNA structures stable. Base stacking is determined by the interaction between the neighboring base pairs. This kind of strong, short range force makes RNA folding non-synergistic, and once a relatively stable structure is formed during the formation of RNA, it is difficult to destroy [25]. Such a structure is in a metastable intermediate state; moreover, the intermediate state is often a local optimum structure [26]. Therefore, we assume that the process of a dynamic RNA secondary structure formation can be divided into several stages and in each stage, the thermodynamically most stable local structure should be formed, and the structure cannot be opened. Following each change of stage, the RNA structure is gradually formed. When the RNA structure cannot be changed, it's native full structure is finally formed. Within a stage, only the helical regions with advantages in thermodynamics may be formed. They compete with each other and will be combined into the thermodynamically optimal local structure at the end of this stage. Competition for helical region bases occurs only within a stage but does not occur between stages.

As shown in Fig. 2, the folding of an RNA from the initial random coil to the full structure goes through intermediate states and H_i denotes the helical region set composed of the helical regions, which have already formed in the structure under the ith $(0 \leq i \leq m)$ state. H_0 presents the initial sequence obviously, $H_0 = \emptyset$ and H_m presents the final secondary structure. Therefore, the transition process of the RNA state is just the transition process of $H_0, H_1, H_2, \dots, H_m$. The structure transforms from the ith state to the $(i + 1)$th state through stage $S_{i+1}, 0 \leq i \leq m - 1$. In the S_{i+1}, the helical regions competing with each other will form the thermodynamically optimal local structure.

2.3.1 Stage Division of RNA Folding

An RNA is a single-stranded nucleotide sequence; thus, forming a double-stranded helical region would inevitably lead to the formation of a ring. To form a helical region, the factors affecting free energy of the RNA are energy released by pairing, stacking

(negative energy, increasing stability) and energy absorbed by closing the loop structure (positive energy, reducing stability). So the energy contribution to the entire RNA for forming the helical region should be the sum of the positive energy and negative energy and the sum is represented by e. Then attach e as another item in each h of set H, so the elements in H can be expressed as $((start, end, length), e)$. The set of such elements is denoted by $H_{current}$. When $e < 0$, the smaller the value is, the more stability can be brought to the current RNA by forming the corresponding helical region, so the possibility of forming the helical region is greater. In other words, the more stable the RNA, the more urgency for formation of a helical region.

At the end of each stage, the helical regions which can be formed are the helical regions most urgent to form at the current stage, and these helical regions are more stable than others according to thermodynamics. So item e of the elements in $H_{current}$ can be seen as the criteria of the urgency degree of helical region formation. So the elements in $H_{current}$ can be ranked as ascending in terms of item e; then the group of helical regions, which are in front of the rank with similar e, are the helical regions most urgent to form. If corresponding item e of element in $H_{current}$ is greater than 0, it means that assembling the helical region to current structure will make the structure unstable, so it will be removed from $H_{current}$. In fact, usually the stability conditions made possible by e correspond to elements in $H_{current}$, which are in front of the rank, and their similarities suggests these helical regions must form at the same time. Then, these helical regions work to compete as they construct the set $H^*_{current}$, $H^*_{current} \subseteq H_{current}$. The item e of the elements in $H^*_{current}$ should be significantly different from the other elements in $H_{current}$. In order to describe the elements belonging to $H^*_{current}$ conveniently, the First Large Energy Difference (FLED) is defined.

The first step is to rank elements in $H_{current}$ in terms of item e in ascending order and mark the free energy items consecutively with e_i, where i is from 1 to $|H_{current}|$. And then calculate the difference d between adjacent e, i.e., $d_i = e_{i+1} - e_i$. Obviously the number of d_i is $|H_{current}| - 1$. We set a parameter α, $0 \leq \alpha \leq 1$. The process of calculating FLED is as follows:

$$j_0 = |H_{current}|, \tag{1}$$

$$D = \{d_{j_m}\}, d_{j_m} = \max_{i=1\ldots j_{m-1}} \{d_i\}, m = 1, 2, \ldots, \tag{2}$$

$$D^* = \{d_{j_m}\}, \alpha \cdot d_{j_m} > d_{j_{m+1}}, \tag{3}$$

$$j_x = \begin{cases} \min\{j_m\}, & d_{j_m} \in D, if\ D^* = \emptyset \\ \min\{j_m\}, & d_{j_m} \in D^*, if\ D^* \neq \emptyset. \end{cases} \tag{4}$$

The initial j_0 can be got by (1) and it is just the number of d_i. The D set is derived by (2) using the method of recurrence and it is composed by d_{j_m} which is the largest one between d_{j_1} to $d_{j_{m-1}}$; therefore, the change of corresponding e is most obvious. Restrict the elements in D further by (3) to get D^*, the elements of which satisfy the condition $\alpha \cdot d_{j_m} > d_{j_{m+1}}$ and so the elements in D^* are ones with a large change compared to the next element. The parameter α in (3) is a scale parameter which measures the degree of change of two adjacent elements. Section 2.2.3 will discuss this degree of change and its effect in detail. At last, j_x can be obtained by (4) and it is the minimum subscript in all the subscripts of elements in D^* (if $D^* \neq \emptyset$). The corresponding d_{j_x} is the FLED; that is, in the current state of RNA folding, j_x elements in the front of the rank of elements in $H_{current}$ are the helical regions urgent to form. These elements constitute the set $H^*_{current}$ and change the state of RNA by competing with each other. If $H_{current} = \emptyset$, the RNA folding is finished.

2.3.2 Search for the Optimal Structure in a Stage

The helical regions in $H^*_{current}$ will compete with each other in each stage. The RNA structure should always be a thermodynamically local optimal structure at the end of each stage; that is, these helical regions reduce mostly free energy during each stage of their structure through competition, fragmentation and recombination. So the final structure in each stage represents the most thermodynamic and stable local structure formed by each particular helical region.

It is possible that the helical regions in $H^*_{current}$ are incompatible, and that the competition will occur among the incompatible helical regions. The helical regions' success in the competition depends on which structure is most stable to keep the maximum length of this helical region, which means it can reduce the most amount of energy. Meanwhile, the failed helical regions in the competition will lose several base pairs. This suggests that the initial set of candidate helical regions does not have to contain all the helical regions, but only the maximum-length helical regions.

The final result of the helical region competition is a thermodynamic optimal local structure formed by these helical regions. So the simple in-depth manipulation can be used to find the optimal structure in all possible structures formed by the helical region competition. Assuming there are k elements in $H^*_{current}$, it is easy to divide them into two parts:

$$H^*_{current} = \{h_i | i = 1, 2, \ldots, m\} \cup \{h_j | j = m+1, m+2, \ldots, k\}$$

$\forall h_n$ is compatible with h_i, $h_n \in H^*_{current}, i \neq n$
$\exists h_n$ is incompatible with h_j, $h_n \in H^*_{current}, j \neq n$

The first m helical regions do not conflict with any other helical regions in $H^*_{current}$, and each of the other $k - m$ helical regions has at least one base which is occupied by other helical regions.

Extend the h_i into the non-maximum helical regions (including the maximum helical region and the empty helical region) in accordance with candidate helical region selection Rules 1-3 and these helical regions compose the set E_i. If the energy of the RNA containing incompatible helical regions is regarded as infinity, the local structure which has the local minimum free energy $\min\{f_{FE}(\cup_{1 \leq i \leq m} h_i \cup \prod_{m+1 \leq i \leq k} E_i)\}$ is the thermodynamically optimal local structure and will be formed at the end of the stage. f_{FE} is the function used to calculate the free energy of a given secondary structure.

Only nested structures which do not contain pseudoknots are considered, which will be covered in the discussion section.

Fig. 3 shows the folding process of a real tRNA (PDB, ID:2DET) predicted by our method. The a-d illustrations show the four states of this RNA according to their order of appearance during the folding process. The regions marked by the curves in Figs. 3a, 3b, and 3c are parts of the helical regions which will win in the helical region competition in the next stage, and they will be formed before the end of next stage. Fig. 3d shows the clover structure eventually formed by our method. Fig. 3e is the MFE structure predicted by mfold. The mfold only correctly predicted two helical regions and the results were quite different from the clover structure which is the structure tRNAs should form.

2.3.3 The Choice of the Parameter

The accuracy of RNA secondary structure prediction is usually measured by sensitivity and PPV [13], which are defined as:

$$\text{Sensitivity} = \text{True positive pairs}/(\text{True positive pairs} + \text{False negative pairs}), \tag{5}$$

$$\text{PPV} = \text{True positive pairs}/(\text{True positive pairs} + \text{False positive pairs}). \tag{6}$$

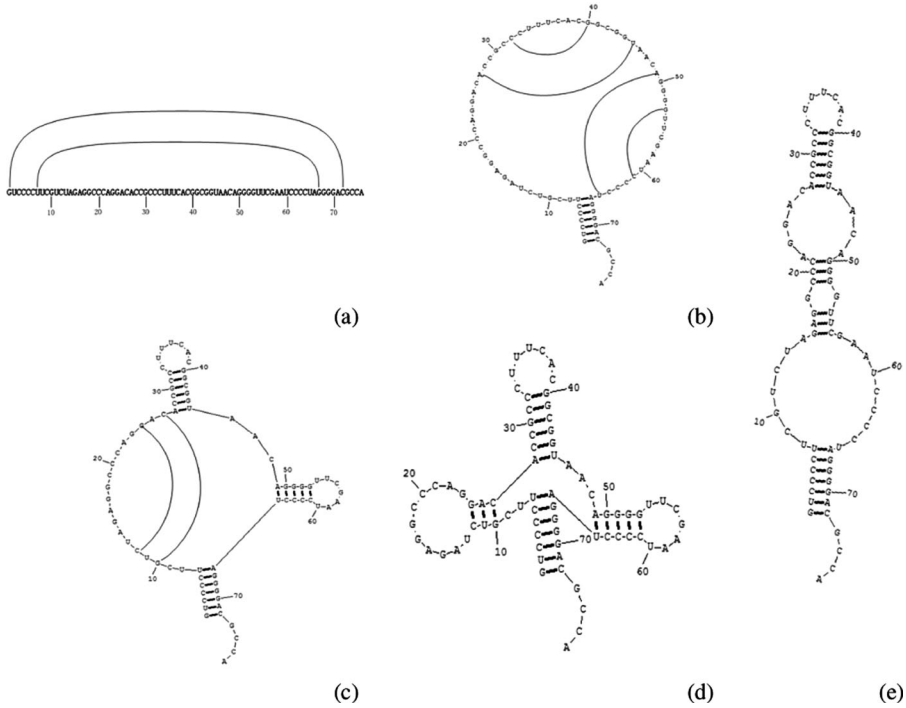

Fig. 3. The folding process of an RNA predicted by fledFold. FledFold is developed to find helical regions most urgently seeking formation among all the possible helical regions.

As shown in (5) and (6), sensitivity measures the ratio of correctly predicted base pairs with native base pairs, and PPV measures the ratio of correctly predicted base pairs using all of the predicted base pairs. These two indexes are widely used in various prediction algorithms.

The selected helical regions at each stage have more advantages than others based on their thermodynamic stability, and parameter α is used to measure to what extent the selected helical regions are better than others. If $\alpha \cdot d_{j_m} > d_{j_{m+1}}$, it shows d_{j_m} was significantly greater than $d_{j_{m+1}}$ so it may become FLED. In all such d_{j_m}, the one with the smallest j_m will become the final FLED. The value of α is between [0,1]. According to (1) to (4), if $\alpha = 0$, there is only one helical region in $H^*_{current}$ at each stage, and this helical region can most reduce the free energy of the current structure. If $\alpha = 1$, all the helical regions before the maximum free energy difference will be selected into $H^*_{current}$.

Different values of α have different effects on the prediction accuracy. Because α is the only parameter in our method, we can test each possible value to select the one which can make our method perform best. We randomly selected 500 sequences (including tRNA, 5S rRNA, RNase P, tmRNA, GI intron and each one of these types included 100 sequences) as test set. We gave α different values between [0,1] every 0.05 and recorded the sensitivity and PPV. Fig. 4 shows the trend of the average sensitivity and average PPV is almost the same; When $\alpha = 0.8$, sensitivity and PPV reached their highest points (sensitivity $= 0.637$, PPV $= 0.624$) at the same time. As α increased to 1, both the sensitivity and PPV dropped down quickly. As the α gradually reduces to zero, both the sensitivity and PPV declined generally except for some individual points. In addition, in the extreme case $\alpha = 0$, our method degenerated into a greedy algorithm [22], and the accuracy in this case was not satisfactory. Therefore, the default value of α is 0.8 in our method.

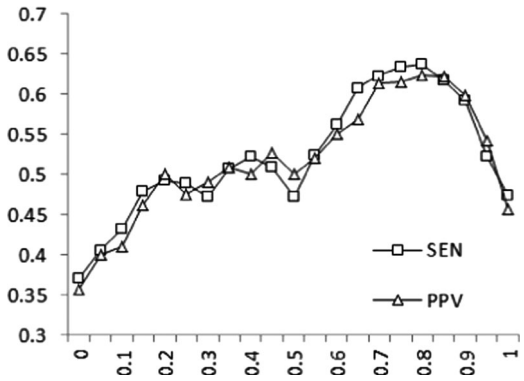

Fig. 4. The relation between the value of α and prediction accuracy.

3 Results

The results of our method are presented in this section and our method will be compared with the RNAfold [15], mfold [14], cofold [20] and Sfold [27]. For RNAfold, mfold and Sfold, we only selected the predicted structures with the minimum free energy.

3.1 Free Energy Model and Date Set

We used the NN model [28] to calculate the free energy of an RNA secondary structure. NN model assumes that the total free energy of an RNA is the sum of individual structural units (helical regions, loops), and neglects the interaction between them. Turner and Mathews [28] in 1999 established the thermodynamic parameters for the NN model most widely used in many RNA secondary structure prediction algorithms, which still hold true for current NN model. Hence, our method also used their 1999 thermodynamic parameters. In order to even out the effect of the thermodynamic parameters on the results, we also used Turner and Mathew's 1999 thermodynamic parameters for mfold, RNAfold, cofold.

All RNA sequences in the data set we used for testing came from tRNA DB 2009 [29], tmRNA Database [30], Nucleic Acid Database [31], RNase P Database [32], and the RNA STRAND Database [33], which are the databases most widely used currently. Testing sequences included tRNA, 5S rRNA, tmRNA, RNase P, GI intron, and the length of them ranged from 70 to 1058. The

TABLE 1
The Comparison on tRNA and 5S rRNA

Method	tRNA		5S rRNA	
	SEN	PPV	SEN	PPV
mfold	0.654	0.609	0.693	0.704
RNAfold	0.660	0.607	0.694	0.704
cofold	0.640	0.596	0.585	0.591
Sfold	0.587	0.650	0.703	0.733
fledFold	**0.707**	**0.708**	**0.715**	**0.743**

The items in bold are the highest sensitivity or PPV among all the algorithms. SEN indicates sensitivity.

TABLE 2
The Comparison on tmRNA, RNase, and GI intron

Method	tmRNA		RNase P		GI intron	
	SEN	PPV	SEN	PPV	SEN	PPV
Mfold	**0.529**	0.486	0.596	**0.594**	0.599	0.504
RNAfold	0.526	0.490	0.594	0.589	0.596	0.500
Cofold	0.474	0.448	0.563	0.569	0.610	0.534
Sfold	0.488	0.524	0.582	0.590	0.576	0.548
fledFold	0.526	**0.536**	**0.617**	0.575	**0.623**	**0.560**

The items in bold have the highest sensitivity or PPV among all the algorithms. SEN indicates sensitivity.

sequences of tRNA, 5S RNA, tmRNA, RNase P and GI intron came from tRNA DB 2009, RNA STRAND, tmRNA Database, RNase P Database, and Nucleic Acid Database, respectively. All secondary structures of these sequences are known and can be viewed as native structures.

3.2 Accuracy Comparison

Tables 1 and 2 compare the results of our method (fledFold) with the other algorithms. The test set of each kind of RNA includes 100 sequences selected randomly from the databases. The value of parameter α of fledFold is 0.8.

Table 1 compares the prediction results of fledFold and other algorithms on tRNA and 5S rRNA. Obviously, the sensitivity and PPV of our method on the tRNA and 5S rRNA are higher than the other algorithms. On tRNA, the sensitivity and PPV of fledFold are 4.7 and 5.8 percent higher, respectively than other algorithms, which perform best. On 5S rRNA, the sensitivity and PPV of fledFold is 1.2 and 1.0 percent higher, respectively, than any other high-performance algorithm. Therefore, our method has more advantages than others on short sequences without pseudoknots.

Table 2 compares the prediction results of fledFold and other algorithms on the longer sequences including tmRNA, RNase P, and GI intron. The accuracy declines for all of the algorithms in predicting these sequences. However, the sensitivity on tmRNA and the PPV on RNase P are 0.3 and 0.9 percent lower than mfold. When predicting tmRNA, fledFold output less true positive pairs and the false positive, but more false negative pairs and when predicting RNase P, fledFold output more false positive pairs but less true positive pairs and false negative pairs. The remaining items in the table shows our method achieving a higher accuracy than the other algorithms, but the accuracy of fledFold is not satisfactory enough. This may be because a long RNA sequence usually folds into a pseudoknoted structure and the pseudoknots may affect the folding state. Our method does not consider the pseudoknots, which may affect the accuracy of the prediction.

4 Discussion

Our RNA structure prediction method takes into account the dynamic process of the RNA folding from random coil to full structure, combines thermodynamics and kinetics, and improves prediction accuracy to some extent. After testing experiments, sensitivities of our method on tRNA, 5S rRNA and GI intron were 4.7, 1.2 and 1.3 percent higher than the best performance of other algorithms respectively; PPVs were 1.9, 1 and 1.2 percent higher than the best performance of other algorithms respectively. The sensitivity on tmRNA and PPV on RNase P of our method was not the best among all the testing algorithms but the gaps were not wide. The PPV on tmRNA and the sensitivity on RNase P of our method was 1.2 and 2.1 percent higher than the best performance of other algorithms, respectively.

The improved prediction accuracy can be due to that the microenvironment where the RNA folds are very different and do not produce traditional paths. Consequently, these differences may cause RNAs to not fold along the fixed paths toward optimal structures. This means that the probabilities which accompany the more common folding paths are not expected to the same for RNAs. Energy traps encountered in the folding process can hinder the formation of an RNA and negatively affect the thermodynamics optimal structure along one or several paths. Nature's RNAs may differ substantially from random sequences in their folding pattern, and would be valuable to understand how they fold into their native structures accurately. However, we are trying to explore the folding pattern of RNAs. Our method does not determine the pathways, but finds the relatively stable key statuses, which avoids the differences on the folding pathway caused by microenvironments.

Our method found all the helical regions complying with the rules in advance based on helical region distribution which greatly reduced the search space and made subsequent calculations simple and fast. Some steps could even be executed concurrently on multiple machines. But these rules may exclude some native helical regions, which could have affected the accuracy of the prediction. So how to restrict candidate helical regions reasonably is another problem that needs to be solved in order to improve our method.

From the program execution point of view: for the longer RNAs, local structures are usually formed by the regions nearest to each other at the beginning of the execution. Along with the formation of such local structures, the helical regions with larger intervals are dragged together and formed. Actually, along with such progress, RNAs transform from their disordered, unstable states to stable, ordered states.

Although the energy parameters Turner and Mathews established in 1999 [28] are widely used, they are still not perfect. As Hudson et al. found, the $\psi \bullet A$ (ψ represents pseudouridine) base pair have significant difference with the traditional $U \bullet A$ base pair on free energy [34]. So energy parameters may also be a limiting factor in the accuracy of our method. At present, energy parameters for pseudoknots have not been experimentally determined [13] and to the best of our knowledge, no good pseudoknot energy models exist. We used the same pseudoknot model in our method as that used for hotknots [35] to predict pseudoknoted structures, but the results were not satisfactory. There are many false negative pseudoknots in predicted structures, which affect the prediction process; subsequently, the prediction accuracy is reduced. Even though our method does not currently consider pseudoknots, we acknowledge that pseudoknots may affect the folding state. Overlooking pseudoknots may affect the accuracy of predictions, especially when predicting the structure of a long RNA sequence. A reliable pseudoknot energy model would further improve the accuracy of our method.

Besides the factor of dynamic folding, other important factors can directly affect the results of RNA folding, such as co-transcriptional folding, transcription pausing, protein-RNA interactions, RNA-ligand interactions etc. These factors are likely to cause the energy of local structures to change, thereby causing a change in the process of folding and sometimes—even causing RNAs to fold into different structures and perform different functions [36], [37].

At present, in our lab's research on drug resistance to gastric cancer, we have used our fledFold method to predict the local structures of 500 lncRNAs differentially expressed in drug resistant gastric cancer cells. Using structural information, we successfully screened the lncRNAs related to drug resistance. Moreover, screened lncRNAs have been confirmed by biological experiments, which also supports the practical value of our method.

Acknowledgments

The authors would like to acknowledge the support of the National Natural Science Foundation of China (NSFC) under Grant No. 61471181, Natural Science Foundation of Jilin Province under Grant No. 20150101056JC, Jilin University postdoctoral research. Hao Zhang is the corresponding author.

References

[1] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 8, 1970.
[2] T. R. Cech, A. J. Zaug, and P. J. Grabowski, "*In Vitro* splicing of the ribosomal rna precursor of tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence," *Cell*, vol. 27, no. 3, Pt 2, pp. 487–496, Dec. 1981.
[3] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz, "The complete atomic structure of the large ribosomal subunit at 2.4 a resolution," *Science*, vol. 289, no. 5481, pp. 905–20, Aug. 11, 2000.
[4] F. Schluenzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath, "Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Angstroms Resolution," *Cell*, vol. 102, no. 5, pp. 615–623, Sep. 1, 2000.
[5] B. T. Wimberly, D. E. Brodersen, W. M. Clemons, Jr., R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan, "Structure of the 30S ribosomal subunit," *Nature*, vol. 407, no. 6802, pp. 327–339, Sep. 21, 2000.
[6] G. Meister, and T. Tuschl, "Mechanisms of gene silencing by double-stranded RNA," *Nature*, vol. 431, no. 7006, pp. 343–349, Sep. 16, 2004.
[7] J. S. Mattick, "Probing the phenomics of noncoding RNA," *Elife*, vol. 2, p. e01968, 2013.
[8] B. Onoa and I. Tinoco, Jr., "RNA Folding and unfolding," *Curr. Opin. Struct. Biol.*, vol. 14, no. 3, pp. 374–379, Jun. 2004.
[9] J. A. Jaeger, D. H. Turner, and M. Zuker, "Improved predictions of secondary structures for RNA," *Proc. Nat. Acad. Sci. USA*, vol. 86, no. 20, pp. 7706–7710, Oct. 1989,
[10] N. R. Pace, B. C. Thomas, and C. R. Woese, "Probing RNA structure, function, and history by comparative analysis," *Cold Spring Harbor Monograph Series*, vol. 37, pp. 113–142, 1999.
[11] R. R. Gutell, J. C. Lee, and J. J. Cannone, "The accuracy of ribosomal RNA comparative structure models," *Curr. Opin. Struct. Biol.*, vol. 12, no. 3, pp. 301–310, Jun. 2002.
[12] D. H. Mathews, "Revolutions in RNA secondary structure prediction," *J. Mol. Biol.*, vol. 359, no. 3, pp. 526–532, Jun. 9, 2006.
[13] M. G. Seetin, and D. H. Mathews, "RNA structure prediction: An overview of methods," *Methods Mol. Biol.*, vol. 905, pp. 99–122, 2012.
[14] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3406–3415, Jul. 1, 2003.
[15] M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic Acids Res.*, vol. 9, no. 1, pp. 133–148, Jan. 10, 1981.
[16] S. Bellaousov, and D. H. Mathews, "Probknot: Fast prediction of RNA secondary structure including pseudoknots," *RNA*, vol. 16, no. 10, pp. 1870–1880, Oct. 2010.
[17] M. Zuker, "On finding all suboptimal foldings of an RNA molecule," *Science*, vol. 244, no. 4900, pp. 48–52, Apr. 7, 1989.
[18] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *J. Mol. Biol.*, vol. 288, no. 5, pp. 911–940, May 21, 1999.
[19] S. R. Morgan and P. G. Higgs, "Evidence for kinetic effects in the folding of large RNA molecules," *J. Chemical Phys.*, vol. 105, no. 16, pp. 7152–7157, 1996.
[20] J. R. Proctor and I. M. Meyer, "Cofold: An RNA Secondary structure prediction method that takes co-transcriptional folding into account," *Nucleic Acids Res.*, vol. 41, no. 9, p. e102, May 2013.
[21] W. Saenger, *Principles of Nucleic Acid Structure*. New York, NY, USA: Academic, 1984.
[22] J. P. Abrahams, M. van den Berg, E. van Batenburg, and C. Pleij, "Prediction of RNA secondary structure, including pseudoknotting, by computer simulation," *Nucleic Acids Res.*, vol. 18, no. 10, pp. 3035–3044, May 25, 1990.
[23] L. WuJu and W. JiaJin, "Prediction of RNA secondary structure based on helical regions distribution," *Bioinf.*, vol. 14, no. 8, pp. 700–706, 1998.
[24] X. Chen, S. M. He, D. Bu, F. Zhang, Z. Wang, R. Chen, and W. Gao, "Flexstem: Improving predictions of RNA secondary structures with pseudoknots by reducing the search space," *Bioinf.*, vol. 24, no. 18, pp. 1994–2001, Sep. 15, 2008.
[25] C. S.-J. T. A. N. Zhi-Jie and C. A. O. S. Z. Wen-Bing, "The statistical mechanics of RNA folding," *Physics*, vol. 3, p. 012, 2006.
[26] P. G. Higgs, "RNA secondary structure: Physical and computational aspects," *Q Rev. Biophys*, vol. 33, no. 3, pp. 199–253, Aug. 2000.
[27] Y. Ding and C. E. Lawrence, "A statistical sampling algorithm for RNA secondary structure prediction," *Nucleic Acids Res.*, vol. 31, no. 24, pp. 7280–301, Dec. 15, 2003.
[28] D. H. Turner, and D. H. Mathews, "NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure," *Nucleic Acids Res.*, vol. 38, no. Database issue, pp. D280–2, Jan. 2010.
[29] F. Juhling, M. Morl, R. K. Hartmann, M. Sprinzl, P. F. Stadler, and J. Putz, "tRNAdb 2009: Compilation of tRNA sequences and tRNA genes," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D159–62, Jan. 2009.
[30] E. S. Andersen, M. A. Rosenblad, N. Larsen, J. C. Westergaard, J. Burks, I. K. Wower, J. Wower, J. Gorodkin, T. Samuelsson, and C. Zwieb, "The tmRDB and SRPDB resources," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D163–D168, Jan. 1, 2006.
[31] H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S. H. Hsieh, A. R. Srinivasan, and B. Schneider, "The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids," *Biophys J.*, vol. 63, no. 3, pp. 751–9, Sep. 1992.
[32] J. W. Brown, "The ribonuclease P database," *Nucleic Acids Res.*, vol. 27, no. 1, p. 314, Jan. 1, 1999.
[33] M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon, "RNA Strand: The RNA secondary structure and statistical analysis database," *BMC Bioinf.*, vol. 9, p. 340, 2008.
[34] G. A. Hudson, R. J. Bloomingdale, and B. M. Znosko, "Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides," *RNA*, vol. 19, no. 11, pp. 1474–82, Nov. 2013.
[35] J. Ren, B. Rastegari, A. Condon, and H. H. Hoos, "Hotknots: Heuristic prediction of RNA secondary structures including pseudoknots," *RNA*, vol. 11, no. 10, pp. 1494–504, Oct. 2005.
[36] R. R. Breaker, "Prospects for riboswitch discovery and analysis," *Mol. Cell*, vol. 43, no. 6, pp. 867–79, Sep. 16, 2011.
[37] B. Chetnani and A. Mondragon, "Structural biology: RNA exerts self-control," *Nature*, vol. 500, no. 7462, pp. 279–80, Aug. 15, 2013.