A Novel Efficient Simulated Annealing Algorithm for the RNA Secondary Structure Predicting with Pseudoknots

Zhang Kai[1,2](✉) and Lv Yulin[1]

[1] School of Computer Science,
Wuhan University of Science and Technology, Wuhan 430081, China
zhangkai@wust.edu.cn
[2] Hubei Province Key Laboratory of Intelligent Information Processing
and Real-Time Industrial System, Wuhan 430081, China

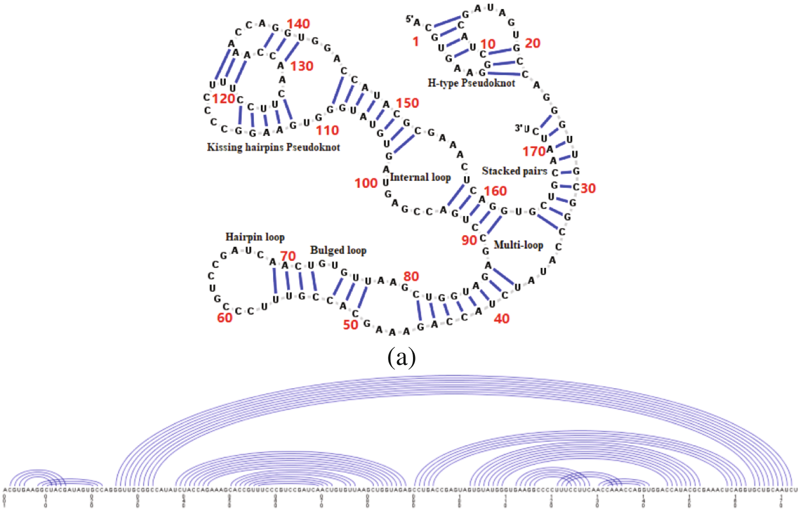
Abstract. The pseudoknot structure of RNA molecular plays an important role in cell function. However, existing algorithms cannot predict pseudoknots structure efficiently. In this paper, we propose a novel simulated annealing algorithm to predict nucleic acid secondary structure with pseudoknots. Firstly, all possible maximum successive complementary base pairs would be identified and maintained. Secondary, the new neighboring state could be generated by choosing one of these successive base pairs randomly. Thirdly, the annealing schedule is selected to systematically decrease the temperature as the algorithm proceeds, the final solution is the structure with minimum free energy. Furthermore, the performance of our algorithm is evaluated by the instances from PseudoBase database, and compared with state-of-the-art algorithms. The comparison results show that our algorithm is more accurate and competitive with higher sensitivity and specificity indicators.

Keywords: RNA secondary structure · Pseudoknot
Simulated annealing algorithm · Minimum free energy

1 Introduction

RNA is a long chain of nucleotides acid molecule which consists of A (Adenine), U (Uracil), G (Guanine) and C (Cytosine). The four-base arrangement allows RNA to have a variety of functions that can play a role in genetic coding, translation, regulation, and gene expression. The search for the secondary structure of RNA sequence has been widely used as the first step in understanding biological functions [1]. The RNA secondary structure folds itself by forming hydrogen bonds between G-C, A-U, and G-U. Therefore, the prediction of RNA secondary structure is returned to predict all hydrogen connections from the primary structure of the sequence. Many components can be identified in the secondary structure, such as stacked pairs or stacks, hairpin loop, multi-branched loop or Multi-loops, bulge loop, and internal loop. The component structures can be represented by a schematic representation or arc representation, as shown in Fig. 1.

D.-S. Huang et al. (Eds.): ICIC 2018, LNCS 10955, pp. 365–370, 2018.
https://doi.org/10.1007/978-3-319-95933-7_44



(a)

(b)

Fig. 1. (a) Typical RNA Secondary Structure. (b) Arc representation of a typical RNA secondary structure. This image was created using jViz.Rna [2].

Pseudoknots usually contains not well-nested base pairs, as shown in Fig. 1(b). These non-nested base pairs make the presence of pseudoknots in RNA sequences more difficult to predict by dynamic programming, which use a recursive scoring system to identify paired stems. The general problem of predicting minimum free energy structures with pseudoknots has been shown to be NP-complete [3].

The dynamic programming (DP) is the first computational approach used to predict RNA structure [4–8]. It can be seen that the temporal and spatial complexity of the prediction algorithm for dynamic programming is high, which is not good for the algorithm to make predictions for long sequence because it will take more time and resources. The other prediction approaches are based on heuristic methods and thermodynamics models [9–13].

In this paper, we propose a novel efficient simulated annealing algorithm to predict nucleic acid secondary structure with pseudoknots. The performance of our algorithm compared with RNA structure method using PseudoBase [14] benchmark instances. The comparison result shows that our algorithm is more accurate and competitive with higher sensitivity and specificity values.

2 Problem Defines

For a given RNA sequence X = 5'$-x_1x_2\ldots x_n-$3' of length n, $M(X)$ is the mapping string of complementary base-pairs of X, $M(X) = (m_1, m_2, \ldots, m_i, \ldots, m_n)$. Each m_i corresponds to the form of (i, j, k), which is called k successive base pairs, where i and j are the base position, where k is the number of successive base pair, and two constraints must be satisfied:

Base Pairs Constraint: If $(i,j,k) \in M$, then $\{(x_i, x_j), (x_{i+1}, x_{j-1}), \ldots, (x_{i+k-1}, x_{j-k+1})\} \in \{(\mathrm{A, U}), (\mathrm{G, C}), (\mathrm{G, U})\}$ in RNA.

K Successive Base Pairs Constraint: If $(i,j,k) \in M$, then $j - i > 2 * MinStem + Minloop$, $MinStem \geq 2$, $Minloop \geq 3$ and $k \geq MinStem$, where *MinStem* is the minimum number of stack and *MinLoop* is the minimum number of loop (Fig. 2). Such as there must be at least *Minloop* unpaired bases in a hairpin loop.

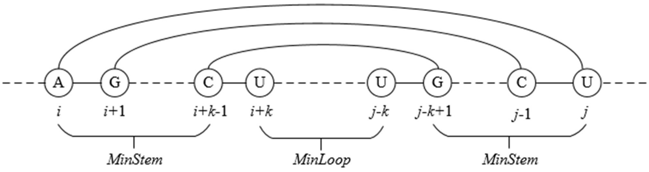


Fig. 2. A graphical illustration of a *MinStem* and *MinLoop*

3 The Proposed Approach

3.1 Set of K Successive Base Pairs

In computer-simulated base pairing, we do not pair individual bases but use successive base pairs. We reduced the range of all possible base pairs by setting *MinStem* and *MinLoop* parameters. Assume that there are three variables i, j, k, which *i* and *j* are the base position, where k is the number of successive base pair. According to the above Fig. 2, we can be seen that i, j, k need to satisfy the following three constraints:

$$1 \leq i \leq RNA.Length - 2 * MinStem - MinLoop + 1 \tag{2}$$

$$i + 2 * MinStem + MinLoop \leq j \leq n \tag{3}$$

$$MinStem \leq k \leq (j - i - MinLoop)/2 \tag{4}$$

3.2 Evaluation Function

For most MFE based RNA secondary structure prediction algorithm, the complex thermodynamic model is often used to evaluate candidate solutions [15]. There are no useful information to guide the candidate solution to find lower neighbor energy state. Consequently, the convergence of these MFE based prediction algorithms is very slow. However, among all of the secondary structure, only the successive base pairs stack structure ΔG_S provide negative free energy which contributes to the reduction of free energy. The stability of RNA sequence can also be approximately evaluated by successive base pairs stacks.

Let $M(X)$ is the mapping string of complementary base pairs of X, $M(X) = (m_1, m_2, \ldots, m_i, \ldots, m_n)$. Each m_i corresponds to the form of (i,j,k), where $m_i\,.k$ equals k, *group* is the number of stems, then the following formula:

$$F(M(X)) = \begin{cases} TotalBP \times AverageBP^2, \boldsymbol{if}\ PesudoGroup < MaxPesudoGroup \\ TotalBP \times AverageBP^2 \times \frac{(TotalGroup - PesudoGroup)}{TotalGroup}, \boldsymbol{else} \end{cases} \tag{5}$$

Where *PseudoknotGroup* is the predicted number of pseudoknot by the algorithm, and *MaxPesudoKnot* is the expected number of pseudoknot.

$$TotalBasePair = \sum\nolimits_{i=1}^{n} m_i.k \tag{6}$$

$$AverageBasePair = TotalBasepair/group \tag{7}$$

3.3 Overall Algorithm

The process of natural RNA folding to its minimal free energy state is very similar to the annealing process. In addition, compared with other heuristic prediction algorithms, such as genetic algorithms, the SA algorithm has faster convergence. Therefore, the paper proposes a new method to predict the RNA secondary structure with pseudoknots based on SA framework. This algorithm framework is as follows:

Algorithm:

```
-Initial Max_T, Min_T, CurrentPairs, MaxPairs.
-While(Temperature>Final_ Temperature) do: //T is current temperature;
    //The upper limit of i is the maximum value of MinLoop
    For (i=0 to RNA.Length-2*MinStem) do
        The new Pair is randomly generated from random set of K successive pairs.
        Remove the conflict match from the CurrentPairs.
        CurrentPairs.Add(Pair);
        ΔE = EnergyDelta(CurrentPairs, MaxPairs, maxPesudoKnot);.
        If(ΔE >=0 OR (Exp(ΔE/T)>Random(0,1)))
            MaxPairs = CurrentPairs;
        End If
    End For
    Decrease Temperature.
-End While.
-Return best solutions: MaxPairs. // MaxPairs is final solutions based on SA.
```

4 Experiments Result

The computational result of our algorithm is compared with IPknot [16], TT2NE [17], CyloFold [18] on 10 benchmark instances in PseudoBase RNA database. The evaluate indicators are *sensitivity* (SN) and *specificity* (SP) [19], as shown in Eq. (8).

$$SN = TP \div TP + FN, SP = TP \div TP + FP \tag{8}$$

Where TP represents the number of correctly predicted base pairs; FP represents the number of incorrectly predicted base pairs; FN represents the number of unpredicted base pairs compared to the known structure. When the prediction results are accurate, both SN and SP should be close to 100%.

The comparisons of the proposed method with the other methods are shown in Table 1. In terms of sensitivity, the proposed method provides the best results in six sequences, yields not the worst result in remaining sequences. In terms of specificity, the proposed method yields the best results in three sequences, similar result in five sequences, and inferior results in two sequences. On average, from all sequences, the proposed method outperforms the other methods in all measure. It has average sensitivity and specificity of 92.6% and 84.3% respectively.

Table 1. Comparison results with sensitivity and specificity indicator

Sequences	[18]	IPknot	TT2NE	OPA	[18]	IPknot	TT2NE	OPA
	Sensitivity				Specificity			
Ec_PK3	85.7	71.4	**100.0**	92.9	**100.0**	76.9	**100.0**	92.9
BEV	93.8	81.3	87.5	**100.0**	**100**	81.3	66.7	76.2
BaEV	86.7	0.0	**100.0**	93.3	**81.3**	0.0	65.2	70.0
VMV	100.0	50.0	92.9	**100**	**73.7**	38.9	65.0	70.0
ALFV	100.0	64.7	**100.0**	**100**	**73.9**	45.8	70.8	70.8
SARS-CoV	69.2	69.2	51.7	**84.6**	72.0	78.3	46.9	**100**
BCRV1	96.7	76.7	**100.0**	**100.0**	85.3	82.1	**96.8**	**96.8**
AMV3	71.8	74.4	74.4	**89.7**	80.0	96.7	72.5	**100**
RSV	97.4	71.8	**97.4**	92.3	88.4	90.3	**90.5**	90.0
CCMV3	66.7	**84.4**	71.1	73.3	66.7	**88.4**	71.1	76.7
Average	86.8	71.5	87.5	**92.6**	82.1	75.4	74.6	**84.3**

5 Conclusion

This paper proposes efficient SA algorithm for the RNA secondary structure predicting with pseudoknots, combined with the evaluation function to compensate for the high time complexity of the free energy calculation model. The algorithm sets the *MinStem* and *MinLoop* parameters to determine the pseudoknot structure formed by the base pair cross-combination, and optimizes the pool of candidate solutions, thereby reducing the time cost of the algorithm. We use the evaluation function to further reduce the time consumption of RNA secondary structure prediction algorithms. Moreover, the performance of our algorithm is compared with state of art algorithms using ten PseudoBase benchmark instances, and the comparison result shows that our algorithm is more accurate and competitive with higher sensitivity and specificity values.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grant No. 61472293). Research Project of Hubei Provincial Department of Education (Grant No. 2016238).

References

1. Jr, T.I., Bustamante, C.: How RNA folds. J. Seq. Biol. **293**(2), 271–281 (1999)
2. Wiese, K.C., Glen, E.: jViz.Rna - an interactive graphical tool for visualizing RNA secondary structure including pseudoknots. In: IEEE Symposium on Computer-Based Medical Systacks, vol. 2006, pp. 659–664. IEEE Computer Society (2006)
3. Wang, C., Schröder, M.S., Hammel, S., Butler, G.: Using RNA-seq for analysis of differential gene expression in fungal species. Methods Mol. Biol. **1361**, 1–40 (2016)
4. Ray, S.S., Pal, S.K.: RNA secondary structure prediction using soft computing. IEEE/ACM Trans. Comput. Biol. Bioinform. **10**(1), 2–17 (2013)
5. Jiwan, A., Singh, S.: A review on RNA pseudoknot structure prediction techniques. In: International Conference on Computing, Electronics and Electrical Technologies, pp. 975–978. IEEE (2012)
6. Rivas, E., Eddy, S.R.: A dynamic programming algorithm for RNA structure prediction including pseudoknots. J. Seq. Biol. **285**(5), 2053–2068 (1999)
7. Reeder, J., Giegerich, R.: Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinform. **5**(1), 104 (2004)
8. Dirks, R.M., Pierce, N.A.: A partition function algorithm for nucleic acid secondary structure including pseudoknots. J. Comput. Chem. **24**(13), 1664–1677 (2003)
9. Ren, J., Rastegari, B., Condon, A., Hoos, H.H.: Hotknots: heuristic prediction of RNA secondary structures including pseudoknots. RNA **11**(10), 1494–1504 (2005)
10. Tsang, H.H., Wiese, K.C.: SARNA-Predict-pk: predicting RNA secondary structures including pseudoknots, pp. 1–8. IEEE (2008)
11. Wiese, K.C., Deschenes, A.A., Hendriks, A.G.: Rnapredict—an evolutionary algorithm for RNA secondary structure prediction. IEEE/ACM Trans. Comput. Biol. Bioinform. **5**(1), 25–41 (2008)
12. Tsang, H.H., Wiese, K.C.: Sarna-predict: accuracy improvement of RNA secondary structure prediction using permutation-based SA. IEEE/ACM Trans. Comput. Biol. Bioinform. **7**(4), 727 (2010)
13. Rastegari, B., Condon, A.: Linear Time Algorithm for Parsing RNA Secondary Structure. In: Casadio, R., Myers, G. (eds.) WABI 2005. LNCS, vol. 3692, pp. 341–352. Springer, Heidelberg (2005). https://doi.org/10.1007/11557067_28
14. PseudoBase. http://www.ekevanbatenburg.nl/PKBASE/PKB.HTML. Accessed 11 Mar 2018
15. Andronescu, M., Aguirrehernández, R., Condon, A., Hoos, H.H.: RNAsoft: a suite of RNA secondary structure prediction and design software tools. Nucleic Acids Res. **31**(13), 3416 (2003)
16. Sato, K., Kato, Y., Hamada, M., Akutsu, T., Asai, K.: IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics **27** (13), i85–i93 (2011)
17. Bon, M., Orland, H.: TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. Nucleic Acids Res. **39**(14), e93–e93 (2011)
18. Bindewald, E., Kluth, T., Shapiro, B.A.: Cylofold: secondary structure prediction including pseudoknots. Nucleic Acids Res. **38**(Web Server issue), 368–372 (2010)
19. Baldi, P., Brunak, S.Y., Andersen, C., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics **16**(5), 412 (2000)