Transactions on Mass-Data Analysis
of Images and Signals
Vol. 10, No 1 (2019) 19-29
© P-ISSN1868-6451, E-ISSN 2509-9353
ISBN 978-3-942952-68-2

ibai Publishing
www.ibai-publishing.org

The Computing Algorithm of Barrier Tree Based on the Basin Hopping Graph in RNA Structure

Zhendong Liu[1,2]*, Gang Li[2], and Patrick Wang[3]

[1]School of Computer Science and Technology, Shandong Jianzhu University Jinan, 250101,China
[2]Department of Biostatistics, University of California, Los Angeles, Los Angeles, 90095,USA
[3]College of Computer and Information Science, Northeastern University, Boston, 02115,USA

Email liuzd2000@126.com, patwang@ieee.org

Abstract. It is an NP-hard problem for prediction of RNA folding structure including pseudoknots, biostatistics is one method of biological data mining, the computing algorithm of RNA structure data is the important in biology. we investigate the RNA pseudoknotted structure based on characteristics of the RNA folding structure , the paper first introduce the Basin Hopping Graph(BHG) as a novel model of RNA folding landscape. Our paper gives the computing algorithm of barrier tree based on the BHG, the experimental results in Rfam13.0 and PseudoBase indicate that the algorithm is more effective. We have improved several types of pseudoknots in RNA folding structure, and analyze their possible transitions between types of pseudoknots.

Keywords: RNA Folding Structures; Pseudoknots; Algorithm; Basin Hopping Graph

1 Introduction

Basin Hopping Graph (BHG) is a novel model of RNA folding structures, computing algorithm of RNA folding structure is the important in biological data mining based on BHG. Biostatistics is one method of data mining, RNA folding is a complicated kinetic process.

Computing algorithm of RNA Structure is the important in biological data mining,biostatistics is one method of data mining.RNA folding is a complicated kinetic process. RNAs are three-dimensional molecules in biological system, which perform a wide range of function. RNA is a key component of moleculars in biological processes. The force of RNA molecules is the set of base pairs, RNA molecules can fold into a three-dimensional structure by forming base pairs of A- U,C-G match,and G-U mismatch, a pseudoknot is two overlapping base pairs, pseudoknots are pairs are known to exist in RNAs[1]. RNA secondary structures prediction is the first step to

*Corresponding author

predict RNA tertiary structures in RNA sequence, RNA tertiary structures is more stable structures, some RNA folding structures are legal. It is very difficult to compute large RNA molecules including pseudoknots. It is NP-hard problem to find an optimal RNA structures. Nussinov had studied the case, where the energy function is minimized when the number of base pairs is maximized, he had designed an algorithm of $O(n^3)$ time complexity to predicting RNA secondary structures[2], Nussinov algorithm can not predict RNA structures with pseudoknots. Algebraic dynamic programming algorithm was proposed by Jens and Robert, it was used to find RNA pseudoknotted structure with simple planar pseudoknots[3], the algorithm takes $O(n^2)$ space complexity and $O(n^4)$ time complexity. The algorithm of finding optimal RNA foldings structure had been firstly known by Michael Zuker[4], Rivas and Eddy had presented Pknots algorithm for predicting RNA pseudoknotted structures based on MFE[5], which time complexity and space comlexity are $O(n^6)$ and $O(n^4)$. The predicting problem of RNA secondary structure including pseudoknots is also NP-complete[6], maximizing the number of stacking pairs allowing pseudoknots in a planar secondary structure makes it NP-hard[7], many researchers seek for approximation algorithms for NP-hard problems. In some mimic RNA structures, pseudoknots are apparently existing[8].The heuristic algorithm for finding RNA structures with pseudoknots has been presented by Ren[9]. Several publications indicates the problem of finding the optimum structure including arbitrary pseudoknots is also NP-hard[10]. People can find the more stable structure including arbitrary pseudoknots if RNA secondary structure is modelled by maximum weighted matching[11]. Some sparse-related techniques have also been applied to RNA folding structures[12-14].

Basin Hopping Graph (BHG) is a novel model of RNA folding structures. Each vertex of the Basin Hopping Graph is a local minimum, which represents the corresponding basin in the structure. Its edges connect basins when the direct transitions between them are ‘energetically favorable’. Edge weights endcode the corresponding saddle heights and thus measure the difficulties of these favorable transitions. BHG can be approximated accurately and efficiently for RNA molecules well in the length range accessible

The barrier tree has two disadvantages: (i) It neglects much of the geometric information of the RNA folding structure because the neighborhood relation between basins is ignored (ii) It is high computational cost makes it unfeasible in for RNA molecules with a more base pairs in length. The BHG can overcome these shortcomings to incorporate additional information of neighborhood.

2 Model of RNA Folding Structures

2.1 Terminology

1. RNA Secondary Structure S: Let S be a set of base pairs such as $s_i.s_j$ is a base pair, base s_i or $s_j \in \{A,C,G,U\}$, $1 \le i \le n$.
2. BHG:Basin Hopping Graph, is a novel model of RNA folding structures.
3. MFE: Minimum Free Energy

4. Stem: the RNA structure closed by base pairs (i, j) and (k, l) ∈S, and (i, j), (i+1, j-1), ..., (k, l) are base pairs, i<k<l<j.
5. Pseudoknot: if si.sj and si'.sj'∈S,i<i'<j<j',or i'< i<j'<j, then the RNA base sequence si…si'…sj sj' composes a pseudoknot.

6.K-Stacking Pairs: In the RNA secondary structure, we use (s_i,s_{i+1},…, s_{i+k};s_{j-k},…,s_{j-1}, s_j) to describe k consecutive stacking pairs (s_i, s_j), (s_{i+1}, s_{j-1}); (s_{i+1}, s_{j-1});(s_{i+2}, s_{j-2}); …… ; (s_{i+k-1} ,s_{j-k+1}), (s_{i+k} ,s_{j-k}).
7. Let S=$s_1s_2…s_n$, s_{ij}=$s_i…s_j$.

3 Predicting Algorithm of RNA Folding Structure

RNA sequence was generated with bases A, C, G, U, it is important for RNA experiments in RNA structure prediction using energy parameters [15,16,17]. We randomly selected the RNA sub-sequences in the Rfam13 and PseudoBase to compute experiments[18,19]. The algorithm can compute RNA nested structure and pseudoknotted structure in RNA sequences. Many experiments in RNA pseudoknotted structures indicated that the algorithm has better predicting accuracy averagely. The algorithm can predict more than 4100 bases of RNA sequences. We have designed effective ways to improve the prediction accuracy for long sequences.

Four experiments in family of PseudoBase can be computed less than 15 seconds with quad-core CPU and 32G memory. The experiments show that accuracy of experiments is valuable, the predicting accuracy outperforms existing algorithms, such as PKNOTS algorithm, MWM algorithm, and ILM algorithm etc[20]. Evolutionary algorithm provide a kind of important method in the RNA structure prediction[21],the structural alignment of RNA is proved to be a useful computational technique for identifying ncRNA[22,23], the efficiency of our algorithm is faster than the other related algorithms in the RNA folding structures and target structures[21,22,23].

4 The Model of BHG in RNA Folding Structures

Lemma 1. If x,y are two local energy minima of RNAfolding structures, there exits a zig-zag path connencting x and y.

Given RNA sequence S, its energy structure L is connected. (see Fig. 1).

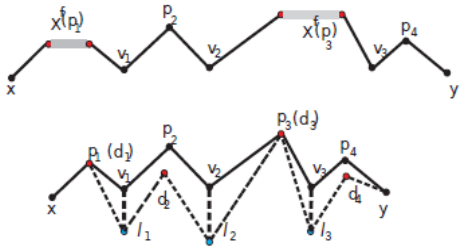

Fig. 1. The path of construction in RNA folding structure

For any substructures given from secondary structures S_1 and S_2, $S_1 \in S, S_2 \in S$, there exists a path between S_1 and S_2, for any two local energy minimum $m1 \in S1, m2 \in S2$, then there exists a zig-zag path,which connecting m_1 and m_2.

We can define the path as follow: path $P=(v_1,v_2,v_3,\ldots,v_k) \in L$, L is the RNA energy structure. If $v_i < v_{i+1}=\ldots= v_{n-1} > v_n$, then for any structures $v_{i+1}=\ldots= v_{n-1}$ are called peak points. If $v_i > v_{i+1}=\ldots= v_{n-1} < v_n$,then for any structures $v_{i+1}=\ldots= v_{n-1}$ are called valley points. If a path P fulfills three conditions:(1) max $f(vk)=S(x,y)$; (2) if $v_i < v_{i+1}=\ldots= v_{n-1} > v_n$,then each vm with $i+1 \le m \le n-1$ is a direct saddle separating the nearest valley points that the path P passed before and after v_m; (3)if $v_i > v_{i+1}=\ldots= v_{n-1} < v_n$,then each vm with $i+1 \le m \le n-1$, there is a minimal shelf L. we declare the path P is a zig-zag path. P can be called Basin Hopping Graph, then Basin Hopping Graph is connected .

RNA structures with pseudoknots can be generalized by the Basin Hopping Graph. We can create sets for implement the gradient walk of RNA structural class with pseudoknots, it comprises 5 types of pseudoknots as follows, Type S, Type H, Type K , Type L Type M. cf. (see Fig. 2). Type S refers to structures without pseudoknots

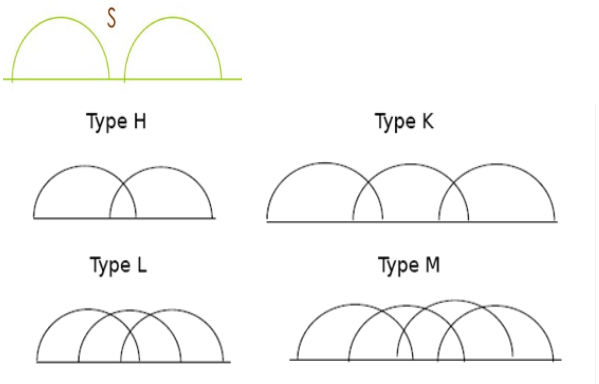

Fig. 2. Types of RNA structural class

5 The Computing Algorithm of Barrier Tree Based on the BHG

Lemma 2. The barrier tree $T_b(V_b, E_b,\omega_b)$ of the RNA folding structure(X,f) is the tree Ta(Va, Ea,ωa) computed by cluster from the complete graph G(V, E,ω).

Vertex sets V_a , V_b and V were including the local minima of the RNA folding structure. Edges sets E_a ,E_b and E weight were including ω({x,y}=S(x,y) for all {x,y}∈E. S(x,y) is the saddle height between any two vertices x∈Va and y∈V_b.

Algorithm Btree T(V, E,ω)
1: M←{(x)|x∈V }
2: V_b←{ (x)|x∈M } and E_b← Ø
3: for all (x)∈V_b do
4: ω_b←f(x)
5: for all (x),(y)∈M×M do
6:$W_{x,y}$←$\omega_{x,y}$ if (x,y) ∈E and $W_{x,y}$←∞ if (x,y) E
7: while |M|＞1 do
8: Search pair of clusters {(u,v)},$W_{u,v}$=min{$W_{x,y}$, (x,y) ∈C(M,2)}
9: for all (x)∈M- {(u),(v) } do
10: $W_{u,x}$= min{ $W_{u,x}$, $W_{v,x}$}
11: create new Tb-vertex(u,v) ←{u}∪ {v} with
ω b(u,v) ←$W_{u,v}$
13: V_b←V_b∪{u,v}
14: E_b←E_b∪{(u,v), (u)} ∪{(u,v), (v)}
15: M ←M-{(v)}
16:End while

The barrier tree Tb can be interpreted into a vertex weighted tree Tb-vertex with the local minima as their leaves. Internal nodes indicate the merging of BHG surrounding two local minima of saddle height. In the step of algorithm Btree, the pairs of clusters are merged which are connected by weight of the smallest edge. Weights of edge are updated to the minimum of the edge weights of the merged clusters in all vertices of separate clusters {(u,v)}. The single linkage clustering implicitly defines a binary tree T, each internal node(u,v)=(u)∪(u) is corresponding to the merging of the clusterings (u) (u), which has the minimuim weight $W_{u,v}$.

The graph B analysed by the algorithm of Btree T and obtain a binary ,vertex-weighted tree, its time complexity is $O(V^3)$.

The algorithm of Btree T has improved in time complexity and space complexity than the other barrier tree Tb(V_b, E_b,ω_b) of the RNA folding structure(X,f) .

Lemma 3. Let BG(V_G,E_G,ω_G) be the Basin Hopping Graph of the RNA folding structure(X, f), V_G denoting the sets of local minima in (X, f), then for all {x,y}∈C(V_G,2), S(x,y)=minmaxω_G({u,v}).

Lemma 4. The barrier tree Tb(Vb, Eb,ωb) of the RNA folding structure(X,f) is the tree Tc(Vc, Ec,ωc) computed by single linkage cluster from the BHG G(V, E,ω).

A Gfold software can implement Boltzmann sampling provided by Reidys. RNA folding structures can be drew by the tool VARNA. the RNA topological structures can be computed. we can study the differences in predicting RNA folding behaviour, and it can be generalized the RNA pseudoknotted framework based on BHG.

There are some examples in possible transitions between types of pseudoknots for RNAfolding strutures in real life.

Removing base pairs is relatively simple since they will never result in an invalid RNA structure, the general case involving five types of pseudoknots is rather involved, even with the restriction to RNA folding structures, with at most one pseudoknot. See Table 1.

Table 1. Possible transitions between types of pseudoknots upon removing a single base pair

Removing	M	L	K	H	S
M	1	1	1	0	0
L	0	1	0	1	0
K	0	0	1	1	1
H	0	0	0	1	1
S	0	0	0	0	1

.

Adding base pairs is also simple since they will never result in an invalid structure, the general case is five types of pseudoknots. See Table 2

Table 2. Possible transitions between types of pseudoknots upon adding a single base pair adding

Adding	M	L	K	H	S
M	1	0	0	0	0
L	1	1	0	0	0
K	1	0	1	0	0
H	0	1	1	1	1
S	0	0	1	1	1

The paper presents an exbmple named PKB92 of tobacco mild green mosaic virus, we investigate 27 bases with pseudoknots named PK1. Its RNA structure can be correctly predicted by the energy of -4.3 kcal/mol by gfold.

(see Fig. 3).

. (((((. [[[[[ ))))).....]]]]] .

Fig. 3. The next pseudoknot-free minimum free energy, RNA secondary is with an energy of 3.9 kcal/mol. (see Fig. 4)..

. (((((.....))))).....

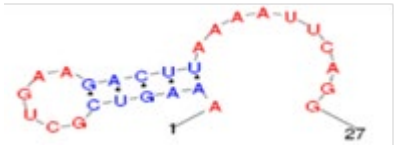

Fig. 4.

It is difficulty that how to determine which base pairs can be added without changing the class of the RNA structure, and compute the changing result in energy without reevaluating the RNA folding structure. We restrict the subset of RNA structures with H-type pseudoknots in restricted class.

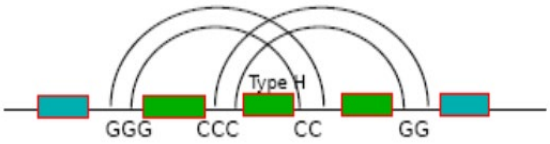

Fig. 5. (A)

An H-type pseudoknots divide the RNA sequence into five regions: two external regions (blue) and three internal regions (green); There are two basic ways to add base pairs. (see Fig. 5).

Fig. 6. (B)

Adding a base pair crossing a stack results in an H-type Pseudoknot in RNA sequences. (see Fig. 6).

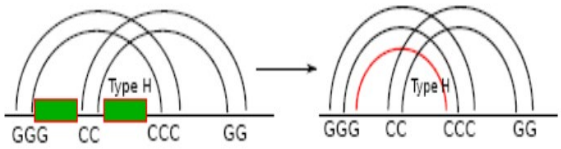

Fig. 7. (C)

Add a base pair which involves two green regions to the existing stacks.
(see Fig. 7).

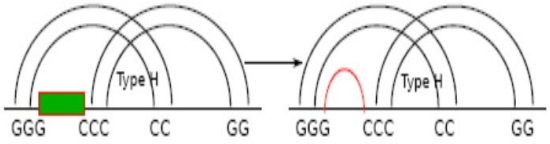

Fig. 8. (D)

Add a base pair which involves nucleotides exactly in one green region without crossing with other existing base pairs. (see Fig. 8).

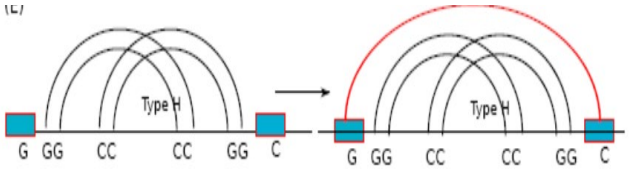

Fig. 9. (E)

Add a base pair which involves two blue regions without crossing other existing base pairs. (see Fig. 9).

According to the principle of the BHG and MFE, the paper provides a path-searching algorithm to connect the graph LM. We investigate the low energy part of the BHG for PKB92 sequence, the PKB92 is more likely to fold the most stable secondary structure, and refold to form the pseudoknots.

6 Conclusion and Future Work

The paper has presented an efficient algorithm for predicting RNA structure with pseudoknots, the predicting accuracy, the time complexity and space complexity outperform existing algorithms, such as MWM algorithm, PKNOTS algorithm and ILM algorithm, Our paper has improved several types of pseudoknots considered in RNA folding structure, and analyze their possible transitions between types of pseudoknots, we also presented the computing algorithm of barrier tree based on the BHG.

It is a also efficient computational method for characterizing the RNA folding structure based on basin hopping graph[24]. The RNA adopts an unexpected tandem three-way junction RNA structure, and unspliced dimeric genomes selected by the RNA conformer may direct packaging[25].

Given an underlying model of gene expression, BayFish uses a Monte Carlo method to estimate the Bayesian posterior probability of the model parameters in smFISH data of single-molecule RNA, RNA Sequencing Reveals and RNA Polymerization in tRNA Fidelity and Repair also are important for RNA structure prediction [26,27].

In the future, we should improve predicting accuracy of the algorithm of RNA folding structures with pseudoknots, and improve the computing algorithm of barrier tree based on the BHG. We should also improve Monte Carlo method to estimate the Bayesian probability of the model parameters in smFISH data of single-molecule RNA in single cells. We also focus on the RNA Sequencing Reveals and their applcation.

Acknowledgements

Our work was supported by the National Natural Science Foundation of China under Grant No.61672328,61672323. We are grateful to T.Akutsu, Lyngsø, Ieong Rivals, and Kucharík for their efficient work.

References

1. R.Nussinov, G.Pieczenik, J.Griggs and D.J.Kleitman.: Algorithms for loop matchings, SIAM J. Appl. Math 35(1), 68- 82 (1978).
2. Michael Zuker.: On finding all suboptimal foldings of an RNA molecule. Science 244, 48-52(1989).
3. E.Rivas and S.R.Eddy.:A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots. Journal of Molecular Biology 285, 2053-2068(1999).
4. Lyngsø and N. S Christian.: Pseudoknots in RNA Pseudoknotted Structure, Proceedings of Recomb. Tokyo Publishing 201-209(2000).
5. S.Ieong, M.Y.Kao, T. W.Lam, et al.: Prediction RNA pseudoknotted structures with arbitrary pseudoknots by maximizing the number of stacking pairs. Journal of Computational Biology 6, 981-995(2003).

6. M. H.Kolk, M.vanderGraff, S.Wijmenga, et al.: NMR structure of a classical pseudoknots interplay of single and double-stranded RNA. Science 280, 434-438(1998).
7. Jihong, Ren, Baharak Rastegari, Anne Condon et al.:HotKnots:Heuristic prediction of RNA pseudoknotted structures including pseudoknots. RNA, 1494-1504(2005).
8. T.Akutsu.: A dynamic programming algorithm for RNA structure prediction with pseudoknots. Discrete Applied Mathematics 104, 45-62(2000).
9. J.E.Tabaska, R.B.Carry, H.N.Gabow and G.D.Stormo.: An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics 14,691-699(1998).
10. Y.Zhang, Y.M.Cheung, B. Xu and W.F. Su.:Detection Copy Number Variants from NGS with Sparse and Smooth Constraints. IEEE/ACM Transactions on Computational Biology and Bioinformatics 14(4), 856-867(2017).
11. R.Backofen, S. D.Tsur, S.Zakov and M.Ziv-Ukelso.: Sparse RNA Folding: Time and Space Efficient Algorithms, Annual Symposium on Combinatorial Pattern Matching, 249-262(2009).
12. R.Backofen, S. D.Tsur, S. Zakov and M.Ziv-Ukelson .: parse RNA Folding: Time and Space Efficient Algorithms. Journal of Discrete Algorithms 9(1), 12-31(2011).
13. D.H.Turner, N.Sugimoto and S.M.Freier.: RNA Structure Prediction. Annual Rewiew of Biophysics Chemistry 17, 167-192(1998).
14. J.A.Jaeger, D. H.Turner and Zuker.: Improved predictions of pseudoknotted structures for RNA. Proc Natl Acad Sci 86, 7706-7710(1989).
15. J.Ruan, G. D.Stormo and W. Zhang.: An Iteratted loop matching approach to the prediction of RNA Pseudoknotted Structures with pseudoknots. Bioinformatics 20, 58-66 (2004).
16. http://www.bio.leidenuniv.nl/~Batenburg/PKBGet.html.
17. Zhendong Liu, Hengwu Li and Damig Zhu.: A Predicting Algorithm of RNA Pseudoknotted Structure Based on Stems. Kybernetes 39(6), 1050-1057(2010).
18. B.Han.: Structural alignment of pseudoknotted RNA. J. Comput. Biol 15, 489-500(2008).
19. Yuping Wang and Chuangyin Dang.:An Evolutionary Algorithm for Global Optimization Based on Level-Set Evolution and Latin Squares. IEEE Transactions on Evolutionary Computation 11(5), 579-595(2007).
20. T.K.Wong.: Structural alignment of RNA with complex pseudoknot structure. J. Comput. Biol. 18, pp.97-108,(2011).
21. Zhendong Liu,Daming Zhu,Hongwei Ma.: Predicting Scheme of RNA folding Structure including Pseudoknots. International Journal of Sensor Networks 16(4), 229-235(2014).
22. Marcel Kucharik, Ivo L. Hofacker,Peter F. Stadler and Jing Qin.:Basin Hopping Graph: A computational framework to Characterize RNA folding landscapes. Bioinformatics 30(14)2009-2017 (2014).
23. C. Sarah, Keane, Xiao Heng, Kun Lu. et al.: Structure of the HIV-1 RNA packaging Signal. Science 348(6237), 917-921(2015).
24. Mariana Gómez-Schiavon, Liang-Fu Chen, Anne E. West and Nicolas E. Buchler.: BayFish: Bayesian inference of transcription dynamics from population snapshots of single-molecule RNA FISH in single cells. Genome Biology 18:164 (12 pages) (2017)
25. Zhendong Liu,Daming Zhu and Qionghai Dai.: Predicting Model and Algorithm in RNA Folding Structure Including Pseudoknots. International Journal of Pattern Recognition and Artificial Intelligence 32(10), (17 pages) (2018).
26. Yury V. Malovichko, Kirill S. Antonets, Anna R. Maslova, Elena A. Andreeva, Sergey G. Inge-Vechtomov and Anton A. Nizhnikov.: RNA Sequencing Reveals Specific Tran-

scriptomic Signatures Distinguishing Effects of the [SWI+] Prion and SWI1 Deletion in Yeast Saccharomyces cerevisiae. Genes 10(3), 212(2019).
27. Allan W. Chen, Malithi I. Jayasinghe, Christina Z. Chung, Bhalchandra S. Rao, Rosan Kenana, Ilka U. Heinemann and Jane E. Jackman.: the Role of 3′ to 5′ Reverse RNA Polymerization in tRNA Fidelity and Repair. Genes 10(3), 250(2019).