RNA SECONDARY STRUCTURE PREDICTION, A CUCKOO SEARCH APPROACH

Amandeep Sharma[1], Amanpal Singh[2]

[1]*Department of Computer Science & Engg., RIEIT Ropar, India,* ***amansharma10408@gmail.com***
[2]*Department of Computer Science & Engg., RIEIT Ropar, India,* ***amanpalrayat@gmail.com***

Abstract

RNA secondary structure prediction uses techniques like crystallography, NMR spectroscopy etc. Computation based techniques estimate the possible base pairs that could be formed in RNA. Soft computing techniques generally select some random pair or pair sequences and then check them according to some parameters. The final sequence of RNA which is closest to the required fitness is selected as the final structure. The cuckoo search approach is good for finding the feasible search space locations. Cuckoo search approach feasibly provides results for the detection of base pairs in the RNA and the RNA secondary structure.

Keywords: *DNA, RNA, base pairs, pseudo-knots, structure, soft computing, techniques.*

----------***----------

1. INTRODUCTION

Soft computing provides cost effectiveness for space or time or both but for a near optimal results. These near optimal results may be tolerated in a large number of cases. In cases where the time required for the achievement of exact results is very high the soft computing could be applied very effectively. Better results for just a small loss of accuracy may be achieved. The conventional non computing techniques are like x-ray crystallography[1] or the Nuclear Magnetic Resonance spectrometry[2] not only have very small applicability but also require a lot of time for the estimation of RNA secondary structures. If we try to increase the applicability for these techniques, then there are severe efficiency losses. These techniques also have to be modified according to the type of RNA to be analysed. Although there are some computing servers which also consider the type of RNA for providing very accurate results, but these estimation servers are also very much faster as compared to conventional techniques. Some of the techniques for the prediction of RNA secondary structure, which provide similar results every time are like Mfold[3], Cofold[4] etc. These techniques are based over either free energy calculation or the estimation offolding locations in given RNA sequence. If pseudo-knots are not to be considered then these techniques takes $O(n^3)$ complexity while if the pseudoknots are to be considered then these would take $O(n^6)$ of time complexity. Also the accuracy of these techniques fall exponentially with increase in the number of nucleotides. The RNA123[5] server uses the calcualtion of free energy for the calculation of RNA secondary structure. The RNAaliashape server uses abstract shapes as basis over which the RNA structure is fit into. RNA123 or RNAaliashape[6] use previously defined shapes or structures in the database for estimation of RNA secondary structures.

The soft computing approaches could also be adapted for finding the RNA structure. These approaches could also be used as a further step for the estimates of RNA structures. In the soft computing approaches the general steps for the estimates are as follows:

1. Generate some random pairs or pair sequences according to some pairing criteria.
2. These generated pairs or the sequences then are tested for some conditions like minimum free energy, foldability or largest sequence length available.
3. The sequences or the base pairs which are best according to the conditions in the previous step are choosen for further analysis.
4. In the final results the sequences with best non overlapping base pairs are selected and the RNA secondary structure is formed.

There are also some approaches in soft computing like tabu search which go completely random but according to some objective condition.

2. SOME SOFT COMPUTING TECHNIQUES FOR RNA SECONDARY STRUCTRE PREDICTION

Soft computing techniques like Genetic and evolutionary algorithms[7], simulated annealing[8], Mfold[3], tabu search[9], Particle swarm optimization (PSO)[10], Ant colony optimization (ACO)[11], Artificial neural networks[12], Fuzzy logic[13] etc have been used for the estimation of RNA secondary structure.

A. GENETIC ALGORITHM FOR RNA STRUCTURE PREDICTION

Genetic algorithm[7] is known to have very accurate results and also large applicability. The genetic algorithm have the steps of replication, mutation and selection. The steps in genetic algorithm for the RNA secondary structure

prediction are :

1. Random base pair sequence generation
2. Elongation of sequences according to some conditions.
3. Random mutations in the sequences of RNA. i.e. random base pairs are added and deleted from the sequence.
4. Checking for the fitness according to some condition.

The above steps are repeated until the maximum iterations have been reached or the maximum output from given fitness conditions stabilises.

B. SIMULATED ANNEALING

Simulated annealing(SA)[8] is the process in which an object which have been subjected to a abrupt changes slowly regains its state. The simulated annealing have also provided good results for the estimation of RNA secondary structures. The general steps taken in the simulated annealing for RNA secondary structure prediction are:

1. Initial generation of random structures.
2. Abrupt changes in base pairs of initial structures.
3. Repair of these abruptly changed structures according to the annealing rates.
4. Decision of RNA sequence selection if appropriate or not.

3. KUCKOO SEARCH FOR RNA SECONDARY STRUCTURE PREDICTION

The general cuckoo search algorithm[14] is good for the detection of feasible search space in given solution state search space. The general cuckoo search algorithm is having following steps:

1. Generate random nests in the solution search space.
2. Check for the feasibility of each of the given nests according some probability or checking function.
3. Generate a number of feasible nests.
4. When the count of these feasible nests increases over certain limit, select some of the best possible solutions
5. The above steps are repeated until the optimization function stabilizes or maximum number of iterations are reached.

4. IMPLEMENTATION OF KUCKOO SEARCH FOR RNA SECONDARY STRUCTURE PREDICTION

Using Kuckoo search for the RNA secondary structure prediction has been done as follows:

1. Each of the base pairs or possible base pairs have been considered as a cuckoo nest.
2. These base pairs are first generated randomly and then tested for detection.
3. If the given base pair is not feasible then is dropped immediately but if feasible, then a given priority value according to Watson crick nearest neighbour parameter[15] is provided to the given base pair string.
4. When no more pairs could be generate in the given fold sequence, then the complete sum of feasibility value is taken into consideration.
5. The sequence formed with maximum priority value is selected as the solution.

The above approach leads to very low time complexity as compare to other algorithms if the RNA sequence is very large. The base pair detection only in the above proposed method also relieves the requirement of consideration for pseudoknots in the given RNA sequence i.e. even without the consideration for the pseudoknots the base pairs from these could also be easily detected in the RNA.

5. RESULT AND CONCLUSION

The given Cuckoo search approach have been tested for the RNA sequences with accession numbers Y08511, U02540, U40258, X54252, X05914, AF034620, L19345, J01415, M27605, X67579. A base pair matching approach have been used instead of the general structure based approach.

Time Complexity: The comparison of relative time complexity of the various approaches as compared to the kuckoo search is as given in the table 1. In table n corresponds to the number of nucleotides.

Table 1. Comparison for non-pseudoknotted structures

Sr. No.	Algorithm	Time Complexity
1	Mfold	$O(n^3)$
2	Dynamic Programming	$O(n^3)$
3	SARNA	$O(n^2)$
4	Genetic Algorithm	$O(n^2)$
5	**Cuckoo Search**	**O(n)**

If the pseudoknots are also to be detected by the algorithms then the required time complexities are as in table 2.

The above table shows that time complexity for constant result algorithms would increase a lot if the pseudoknots have also to be detected by the algorithms. The time complexities for the algorithms could be more elaborated as if we take *g* as number of generations for cuckoo, genetic or SA, *h* as number of possible helices to be detected in genetic or SA algorithms, *n*

Table 2. Comparison for pseudoknotted structures

Sr. No.	Algorithm	Time Complexity
1	Mfold	N. A.
2	Dynamic Programming	$O(n^6)$
3	SARNA	$O(n^2)$
4	Genetic Algorithm	$O(n^2)$
5	**Cuckoo Search**	**O(n)**

being the number of nucleotides and *x* being the initial population size. The time complexities for the algorithms would be:

1. For Genetic algorithm the time complexity is *O(gxn)*[16].
2. For SA the time complexity is given as *O(T(h+h+n))*[16]. Where T is the number of time steps taken while annealing or getting into some feasible shape.
3. Here in the base pair Cuckoo search the time complexity comes out to be *O(gx).*

The time complexity of the base pair based cuckoo search is just dependant over the possible base pairs in the given RNA sequence, as:

No. of Base Pairs<=n/2 (2)

So, the time complexity for the approach used here is constrained by *O(n)* only. Making this base pair based cuckoo search as an acceptable choice if we want to go for reducing the estimation time for the large RNA sequences.

Sensitivity: It is a measure used for testing the effectiveness of RNA secondary structure prediction algorithms. The sensitivity is the ratio of base pairs detected by a given algorithm to the ratio of actually present base pairs in given RNA. It is given as :

$$\frac{\text{True Positive base pairs}}{\text{(True Positive + False Negative) base pairs}} \quad (2)$$

The Cuckoo search base pair approach have lead to some unreliable sensitivity results. As such it may not be good for the detection of RNA secondary structure prediction, but much more stable results may be provided by the algorithm if used in hybrid with other algorithms. The direct base pair detection approach leads to easy implementation with other algorithms for the RNA secondary structure prediction.

Specificity: It is defined as the ratio of number of non base pairing RNA nucleotides which have been estimated correctly by the algorithm to actual non pairing nucleotides. The specificity value for the RNA secondary structure prediction algorithm could be given as:

$$\frac{\text{Detected non pairing nucleotides}}{\text{Actual non pairing nucleotides}} \quad (2)$$

The Cuckoo search base pair approach have large result variations for the specificity too but the results on an average are better than the sensitivity results. Table(3) shows the comparison of given algorithm for with SARNA predict specificity results.

Table 3. Comparison with SARNA

Organism	Acession Number	Specificity : SARNA	Specificity: Cockoo Search
S. crevisiae	X67579	84.6	74.8
H. marismortui	AF034620	90	80.2
A. griffini	U02540	51.8	58.6
H. rubra	L19345	48.8	59.3
D. virillis	X05914	33.5	57.9
Homo sapiens	J01415	47.5	66.2

Other features: The approach may be having large sensitivity fluctuations but the single RNA sequences which are having large probability of making base pairs sequences. As the major applications of prediction of RNA secondary structure, are based over RNA sequences. This approach could be useful for the detection of riboswitch locations in the RNA or for better drug discovery of RNA based drugs. If some random cuckoo base pair nests could be neglected then this approach also leads to a very high specificity(another measure for RNA secondary structure prediction) but detected base pair reduction would be very large for this case.

CONCLUSION

This approach have used the detection of base pairs according to Cuckoo search and free energy calculations. Which have lead to very low time complexities for the algorithm. Making it an acceptable approach for RNA structure prediction. But, the approach does not fits into the metrics of structure prediction for RNA folding like specificity or sensitivity. Also the results provided by the given algorithm are good for the general applications of RNA structure prediction like drug detection or the riboswitch location detection(although not better for the applications like RNA classification or gene classification as these are structure based applications). If there are more application based metrics for RNA structure prediction then the results from the given approach could be better. The approach is also good for making hybrids with other algorithms like genetic algorithm or the SA algorithm which require generation of initial structures or have to periodically deform the RNA structure to achieve better structure.

REFERENCES

[1] S.H. Kim, G. Quigley, F.L. Suddath, and A. Rich, “High-Resolution X-Ray Diffraction Patterns of Crystalline Transfer RNA that Show Helical Regions,” Proc. Nat’l Academy of Sciences USA, vol. 68, pp. 841-845, 1971.

[2] A.E. Ferentz and G. Wagner, “NMR Spectroscopy: A Multifaceted Approach to Macromolecular Structure,” Quarterly Rev. of Biophysics, vol. 33, pp. 29-65, 2000.

[3] M. Zuker, D. H. Mathews, D. H. Turner, RNA Secondary Structure Prediction. Current Protocols in Nuclear Acid Chemistry 2007.

[4] J. R. Proctor and I. M. Meyer “CoFold: an RNA

secondary structure prediction method that takes co-transcriptional folding into account" Nucleic Acids Res. 2013 May; 41(9): e102.
[5] dnasoftware, http://www.rna123.com as on 1stNov 2015
[6] S. Jenssen and R. Giegerich "The RNA shapes studio" Oxford Journals, Bioinformatics, Vol 31, Issue 3, Pp 423-425, 2015.
[7]F.H. Van Batenburg, A.P. Gultyaev, and C.W. Pleij, "An APLProgrammed Genetic Algorithm for the Prediction of RNA Secondary Structure," J. Theoretical Biology, vol. 174, no. 3,pp. 269-280, 1995.
[8] Acceleration based Particle Swarm Optimization (APSO) for RNA Secondary Structure Prediction, J. Agrawal, S Agrawal - Progress in Systems Engineering, 2015 – Springer, Volume 330 of the series 'Advances in Intelligent Systems and Computing' pp 741-746
[9] Y. Liu, J. Hao, and J. Peng, "Predicting RNA Secondary Structurewith Tabu Search," Proc. IEEE Int'l Conf. Cognitive Informatics,pp. 409-414, 2010.
[10] Acceleration based Particle Swarm Optimization (APSO) for RNA Secondary Structure Prediction, J. Agrawal, S Agrawal - Progress in Systems Engineering, 2015 – Springer.
[11] N. McMillan, "Rna Secondary Structure Prediction Using Ant Colony Optimisation," master's thesis, School of Informatics,Univ. of Edinburgh, pp. 1-63, 2006.
[12] D.R. Koessler, D.J. Knisley, J. Knisley, and T. Haynes, "A Predictive Model for Secondary RNA Structure Using Graph Theory and a Neural Network," BMC Bioinformatics, vol. 11,pp. S6-S21, 2010.
[13]D. Song and Z. Deng, "A Fuzzy Dynamic Programming Approach to Predict RNA Secondary Structure," Proc. Sixth Int'l Conf. Algorithms in Bioinformatics, pp. 242-251, 2006.
[14]Xin-She Yang, Nature-Inspired Metaheuristic Algorithms, Second Edition, Luniver Press, (2010).
[15] Turner 2004 GU nearest neighbor parameter and Watson-Crick Parameters.
[16] RNA Secondary Structure Prediction Using Soft Computing Shubhra Sankar Ray and Sankar K. Pal, IEEE/ACM transactions on computational biology and bioinformatics,vol.10,no.1, 2013.