International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.076

Prediction of RNA Structure Using Genetic Approach

Jambi Ratna Raja Kumar
Associate Professor, Department of Computer Engineering , GenbaSopanraoMoze College of Engineering,
Balewadi, Pune, Maharashtra, India,
ratnaraj.jambi@gmail.com

S.Jayalakshmi
Associate Professor, Department of Computer Science and Engineering, Veltech Multi Tech DrRangarajan DrSakunthala Engineering College,
Chennai, Tamil Nadu 600062
,jayalakshmi.research@gmail.com

NamineniGireesh
Professor and Head, Department of ECE, Mohan Babu University(Erstwhile SreeVidyanikethan Engineering College),
Tirupati,India, naminenigireesh@gmail.com

Karimulla Syed
Department of Mechanical Engineering, KoneruLakshmaiah Education Foundation,
Guntur, India-522502,
syedkarimulla@kluniversity.in

DurgaprasadGangodkar
Professor, Department of Computer Science & Engineering, Graphic Era Deemed to be University,
Dehradun,Uttarakhand,India,
dr.gangodkar@geu.ac.in

Mohammad Haider Syed
Assistant Professor, Department of Computer Science, College of Computing and Informatics, Saudi Electronic University,
Saudi Arabia,
m.haider@seu.edu.sa

Abstract—The challenge of predicting RNA construction with pseudoknots is NP-complete, and the objective is to achieve the best RNA configuration with the least quantity of electricity. Numerous approaches for predicting RNA frameworks, including pseudoknots have been developed throughout the years. Metaheuristic techniques are influential in determining lengthy RNA frameworks in a small amount of time. For a forecasting RNA secondary structure with pseudoknots, we employed two optimization methods: Optimization Algorithm (GA) and Simulations Annealing (SA). We've also employed a hybrid of these different techniques called GA-SA, in which GA is used for universal searches, but SA was employed to searches, as well as GA-SA, in which SA would be employed of universal searches or GA was employed to organic investigation. The efficiency of such RNA structure was calculated using four main computational methods. Methods were built using five databases derived from the RNA STRAND or Pseudobase++ databases. The algorithms' values are compared to that of various other optimization techniques. On all databases, the conjunction of GA and SA (GA-SA) techniques, as well as the other four state-of-the-art techniques.

Keywords—Simulated Annealing; RNA framework; metaheuristic methods; Genetic Approach

I. INTRODUCTION

All living cells contain ribonucleic acid (RNA), which is an important biomaterial. Transcription and translation are the fundamental features of RNA [1]. RNA polymers play a variety of additional important roles in biological activities, including transporting genetic data, regulating gene transcription, and acting as catalysts [2-4]. It is vital to determine the configurations of RNA to comprehend its operations. Physical technologies of estimating RNA architecture, such as X-ray crystallographic NMR, is costly and time [5]. pseudoknot was the RNA secondary structure in which is stem's unmatched leading nucleotides are coupled with the stem's unbalanced inward region.

Two optimization and famous metaheuristic algorithms are Genetic algorithms and Evolutionary Computation [6]. To benefit the local search heuristic, the simulated annealing algorithm was introduced [7]. It could be used as a metaheuristic for both local and global searches. In most cases, evolutionary algorithms are employed in the search strategy. Numerous studies have employed simulated annealing or evolutionary computation to forecast RNA pseudoknotted structure. The knot method depends on the Genetic Algorithm [8]. Itcreate an x n matrix to represent an RNA secondary framework of duration n. Several rows or columns as in incoming RNA sequence were used to designate essential nucleotide [9]. An I nucleotide and j nucleotide, for instance, is represented by row I and column j, respectively. This same value matrix[i, j] = 1 if a member in the matrix is a conventional Watson-Crick base couple or Wobble base couple (GU); otherwise, matrix[i, j] = 0. After the matrix has been filled, a list of the greatest stems, known as the stems list, is generated [10]. The structure and composition of RNA are then built using various combinations of the maximum stems. An optimization algorithm was then used to determine the best solution with the least amount of free power.

For the adaptation calculations, two alternative thermodynamics, power systems were used. Itemployed the same power spectral density in the alternative stem strongly indicates [11]. Itused the updated (D&P) electronic model to updated attributed as an optimization process to evaluate the fitness for individuals in the genetic method. For confirmation of GAknot, itused two databases. The PK168 database, which was acquired from, features 168 RNA pseudoknotted molecules. Another is HK41, which contains 41 elements and is made up of a subgroup of the frequencies used in HotKnots [12]. The main benefits of their technique are that ithave removed the constraint on single development or have made certain changes to the contractors and chromosome structure to boost reliability.II. RELATED WORKS

Another methodology that enhanced RNA secondary structure prediction employing pseudoknots used a customized version of the free energy function. The DP09 estimation algorithm is used to develop a new proposed scheme. Itbegin by selecting 1057 RNA investment framework pseudoknots from the RNA STRAND database [13] as the training sample. The branches of each chain are then extracted from the learning algorithm and scanned to produce lists of 1-meters, 2-meters, or 3-meters, correspondingly. K-meter was subsequence of such RNA

DOI: rp-9788770040723.076

classification which is k segments long. Following that, itcalculate the counts of each k-mere sequence. The optimization algorithm or the GRASP approach is used to create an approach for predicting RNA secondary framework. The GRASP technique's biggest benefit is that it incorporates the benefits of the search algorithm, neighborhood approaches, and excessive heuristics. Itcalculated the free electricity of the RNA secondary configuration using the Turner framework since 2004. The findings demonstrate that the technique outperforms the other approaches [14]. The biggest drawback is that they've only performed with smaller patterns, and the best option isn't always confirmed.

An approach to forecasting RNA secondaryconfiguration of pseudoknots predicated on evolutionary algorithms. Initially, the largest number of consecutive complimentary base couples is determined. K consecutive base pairs are complementary sequence couples of the form (i, j, k), when i or j were the component locations and k was the amount of consecutive basic couples. The requirement [15] should be met by the base pairs. New nearby nations are created at random using subsequent nucleotide sequences. The annealing system's scheduling attributes are created to gradually reduce the temperature until the RNA structure was solved with the least amount of available electricity. The methodology does not use a thermodynamic framework; instead, it uses sequential basepair stacks to estimate the free electricity of the RNA structure [16-17]. Employed 10 RNA pseudoknotted transcripts from the PseudoBase collection to test the method's capability. The computation median Responsiveness and PPV are 92.6 and 84.3, correspondingly, according to the outcomes.

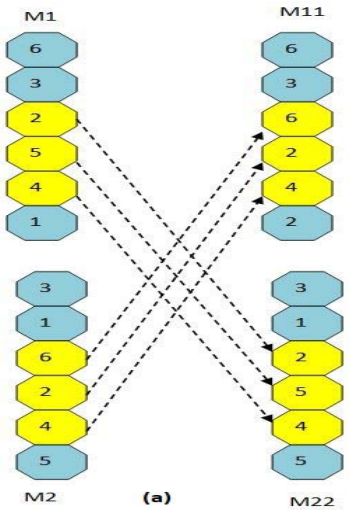
III. Proposed Methods

We used the Simulated Annealing and Genetic Algorithm to estimate RNA secondary system, pseudoknots, in this study. Italso employed the Evolutionary Algorithms but also Simulated Annealing to create two hybrid technologies. GA-SA or SA-GA were the names to the techniques. GA is employed for worldwide searches, whereas SA should be used for local searches in GA-SA. In the SA-GA system, on either extreme, GA was employed for global search, and SA was employed of exploration strategy. There are three basic stages in the methodologies: initiation and population development, iteration, and termination or assessment are all steps in the process.

An RNA pseudoknotted sequence of length n is fed into the population process. Then an unfilled collection named a board of size n n is created. Both the column or row is classified, with the denoted containing the sequence's is a nucleotide. The panel is then covered with the numbers of the set, v = 0 and 1. A stem integer is selected at random from the original population. For each index in the branch quantity, a stem could be taken from the stemmed list. Assume that a stem is expressed (p, q, l). While p denotes to stem's beginning location and q denotes the stem's ending location. In the matrix, p corresponds to I and q corresponds to j. A person is selected at random from the community. The input signal is 21 characters long. As a result, a list is formed, with all integers filled with a dot. From index 19 backward, five opening parenthesis is provided for the stem. The stems of stems 1, 5, and 2 are overlapping. As a result, the architecture is unaffected by these stems. Several alternative possibilities are available for branch 3. First, because the stems are overlapping, branch 3 cannot be inserted. Second, for stem optimization, we can substitute stem 4 with stem 3. Finally, compute the power of both intersecting stems and keep the one with lower power. This method might also be used with the other branches. A helix with fewer seems, on the other hand, generally has more favorable free electricity.

The Genetic Algorithm (GA) is a search-based improvement approach that focuses on biological and ecological choice concepts. It's a well-known approach for finding perfect and nearly-optimal responses to NP-hard temporal making this change. To find the best responses, itused three GA technicians: overlap, evolution, and survivor's choice.

Biological confluence is comparable to the mutation operation. More than one parent is chosen in this function, and one or more offspring are generated. We chose two parents in our method and have them cross to generate two new offspring. Crossover can be carried out in a single location, multiple points, or uniformly. The consistent overlap could undermine the favorable solutions if a single mutation occurs produces a practically identical solution. Itchose a two-point crossover predicated on our retraining research. Two chromosomic, m1 to m2, were randomly selected for community, as seen in



DOI: rp-9788770040723.076

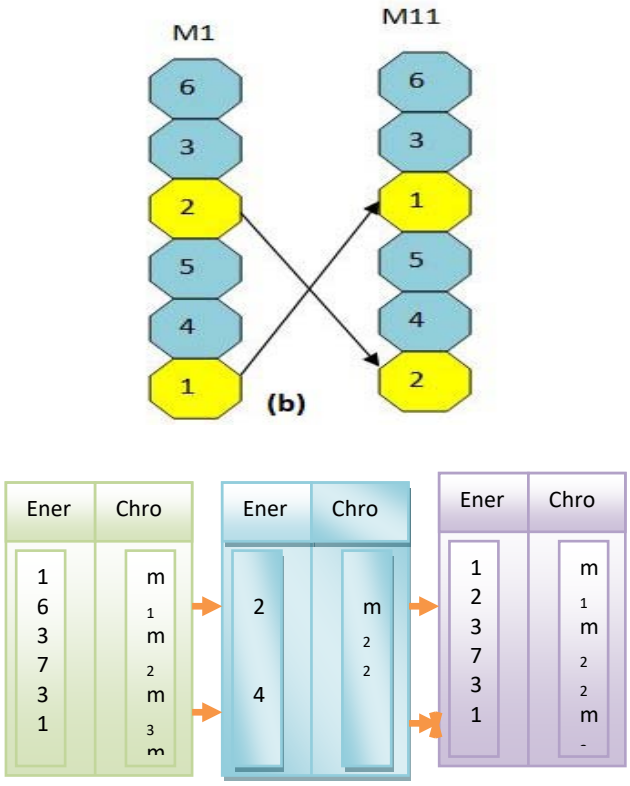


Fig.1.GA 3 Techniques (a) Crossover, (b) Mutation and (c) Selection

Figure. 1(a). To determine the crossovers locations that break each chromosome's three segments, two random digits x and y in the range m1 or m12 were created.For this study, an alternative database called DCC06 to 20 RNA pseudoknotted sequences isselected. The coefficient of Determination improved from repetition 80 to 200. There was no discernible improvement in outcomes after 200 (up to 1000) repetitions. As a result, iteration = 200 is chosen as the optimum number. According to the findings of the experiments, PopSize = 70 is the best fit for achieving improvement. A tiny population may lack all branches, whereas a huge population may have double chromosomesMutation Rate set between 0.05 and 0.4, and the value Mutation Rate = 0.1 yielded the best results. CrossoverRate was also tested from 1.0 to 0.65 before being set at 0.85. A final assessment was conducted using all of these attribute values. It obtains the following result: F-measure = 87.34.

IV. Implementations

The configuration phase utilizes information to Primarily consist or generate the first community. The optimal solutions of the chromosomes are determined at the start of the repetition, and the fittest chromosomes are chosen. The identified community is then subjected to evolutionary algorithms. If no crossover occurs, the offspring is a carbon duplicate of the parent. If it is a crossover, the kids are required up of both fathers' chromosomes. During the crossing, there is no transformation, the progeny is collected without alteration. When a mutation is executed, a portion of the chromosome is altered. CrossoverRate and Mutation Rate are two variables that specify how frequently crossover and mutation will occur. The performance index of the new demographic is computed after each repetition, and the best members are chosen. Variants continue until the halting criteria are satisfied.

The technique of heating systems, copper to modify its internal structure is known as tempering. The metal's new framework is gripped when it cools, and the alloy preserves its newly acquired qualities. Simulated Annealing is a software method that imitates this natural event (SA). The temperature is maintained varied during the crystallization process. The heat is set very high at first, then gradually decreased as the process progressed.

Characteristics employed in SA are listed in Table 1. A variable particular method is summarized in Table 2. The same database that was employed in GA is being used in the procedure as well. This same effectiveness was F-measure = 74.7 when the weather (T), short (T 200). A coefficient of Determination is increasing for temperatures between 200 and 500 degrees Fahrenheit. There was no notable improvement in outcomes after 500 repetitions, and the F-measure was set to 91.3. As a result, T = 500 has been chosen as the ideal temperature range. According to the findings of such tests, PopSize = 70 is an optimal value for maximizing performance. Figure 2 depicts the parameter estimation curve for SA.

TABLE 1.RNA structure SA characteristics

Symbols	Description
T	**Intital temperature of the sytem that is decreased over time**
PopSize	**The population size or the no.of individules in the population space**

TABLE 2. Predicting the RNA Framework

Serial No.	Temperature Celsius	F
I	250	73.8
II	350	83.7
III	450	90.8
IV	1000	90.8

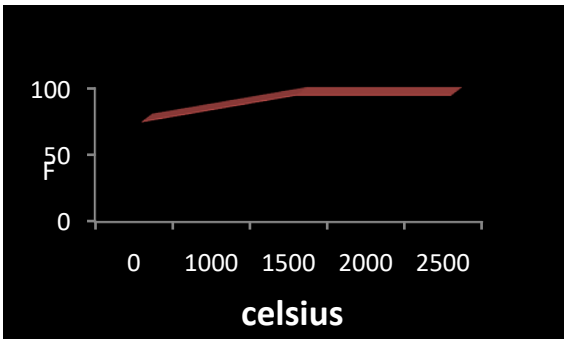


Fig.2.SA parameter modification

Figure 2: SA variable modification The values are defined numbers at the start, but the first community is created. The majority's initial energy is determined, and the person with the lowest energy is chosen. Each cycle generates a set of that individual's neighbors. The person with the minimum energy is chosen from the group of neighbors. For RNA secondary structural characterization to pseudoknots, designers merged to create GA-SA, a hybrid

approach. On the result of GA, we employed SA of exploration, strategy or SA of explorationengine in this methodology. The GA-SA algorithm was run using the same variables as the GA and SA methodologies separately. GA and SA techniques have some drawbacks in this strategy.

As a consequence, we're considering integrating SA-GA to re-search the optimum outcome to shorter stem portions. Another drawback of GA is that it is prone to become caught in a local optimum. Due to the negative influences of its employees, GA occasionally loses better alternatives. An alternate careful evaluation, such as the Gaussian distribution, could be used to compute the likelihood of refusing or adopting a new construction. A minimum guaranteed position could also be found by searching in varying configurations of a person. This could be accomplished by hybridization techniques such as GA GA or SA SA. The analysis revealed that a hybrid GA-SA system outperforms a single GA and SA system.

V. Results and Discussions

To test our methods, we selected five different data sources. The variable sample was chosen with D-CC06, and the test was conducted with another four. The datasets employed in CC06, Evolutionary Computation, and Simulated Annealing are denoted in the chart as D-CC06, D-SA, and D-GA, correspondingly. IPknot yielded the PK168 database, which comprises 168 pseudoknotted RNA sequences, and the HK41 set of data, which includes 41 sequences. On Windows, the techniques were written in Python. For each sequence of the suggested technique, itaveraged the results often runs. Awareness, Positive Predictive Value (PPV), F-measure, and Protein Interaction Integrity are all measures of prediction accuracy (INF). Reflectivity is characterized as the measure of positive examples of the overall amount of positives in the actual course. In a prediction category, the PPV is roughly proportional positives to total positives. The F-measure, which is a proportional evaluation of excitability and PPV, is the weighted harmonic estimate of resonance or PPV. Matthews' coefficient of determination of connection prediction, often known as INF, is a measure of how accurate a prediction is. The following are the numerical expressions for exposed, PPV, F-measure, or INF:

Basepairs are considered positive in RNA structure, while free bases are considered negative. As a consequence, anticipating a correct base pair is TP, whereas failing to forecast a base pair is FN. TN denotes omitting away free bases as such, whereas FP denotes transferring them as a specific gene. Designers did not use TN in RNA-PSPP because it is not specified formulae.

TABLE 3. Comparison performance measures

Dataset	GA	SA	GASA	SA GA	Result of the algorithm mentioned in references
PK168	90.54	80.12	83.47	81.76	74.10,73.4,71.5
HK41	75.65	77.55	88.52	98.22	76.1,76.23,56.75,67.01
DGA	88.66	88.55	40.58	95.25	80.87

TABLE 4. Comparison of proposed with existing system

Dataset	GA	SA	GASA	SA GA	Result of the algorithm mentioned in references
PK168	100.54	82.12	83.47	86.6	76.88,76.5,74.6
HK41	77.55	70.145	84.51	97.44	75.6,7.30,54.50,76.51
DGA	93.47	87.50	77.80	98.60	85.87

Table 3 related that GA-SA would be the greatest PPV in the D-SAor PK168 databases solely, and the HK41 database includes the SA-GA method. On the D-GA database, the technique has the greatest PPV. Furthermore, the PPV of all four of our methods is greater than the PPV of the methodology database. The GA-SA method has a great outcome on all databases, as shown in Tables 4. In addition, SA-GA, along with GA-SA, has a high score on the DGA database. In comparison to previously created algorithms, four of methodologies would be a greater F value or INF. In respect of F value as well as INF, GA-SA method outperforms the database technique.

These optimization methods, SA-GA or GA-SA, exceeded both pure GA but also SA techniques as well as connected comparative methodologies. Designers talked about the drawbacks of using GA or SA to establish our technique, as well as the reasons for and benefits of convergence. For the choice of stems, we employed a variable called u. It aids in the reduction of huge sequence processing time. Itsplit larger RNA architecture branches into numerous smaller stems during hybridization. As a result, the low-energy branches are more thoroughly investigated. The performance of the two methods was significantly improved as a result. We've also used a large data set to refine the methods. That was extremely useful in determining the optimal number of criteria for the algorithms. As an outcome, the technique gives better outcomes when such deserve the right are used. For its smart optimization method, decomposition of huge branches, and full energy estimates, GA-proposed SA's hybrid metaheuristic methodology attained the highest results.

VI. Conclusions

This paper describes an implementation of the Evolutionary methodologies or Simulated Annealing to RNA secondary configuration determination including pseudoknots. The construction to SA-GA or GA-SA optimization methods of tackle the issues is our vital contributor to this research. One of the most difficult jobs was implementing GA to the persons established by SA and implementing SA to GA persons. To determine the optimum RNA structure, four energy estimates were used. The energy calculations for pseudoknotted and pseudoknot fewer structures have been established. In the hybrid version, the local search algorithm looks for all feasible structures based on the results of the universal query optimizer. Almost every type of pseudoknot may be found in our database. It can anticipate pseudoknots of the second element. It cannot, nevertheless, anticipate exceedingly complicated pseudoknots. The findings show that GASA and SA-GA perform significantly better than solo GA or SA outcomes. The GA-SA and SA-GA techniques outperform established methods. The GA-SA algorithm predicts RNA structures

with both short and long sequences quite well. On all databases, GA-SA outperforms three techniques and all four state-of-the-art technologies. In the future, these techniques could be extended to evaluate performance using alternative energy assumptions.

REFERENCES

[1] R. Lorenz, and P.F. Stadler,"RNA secondary structures with limited base pair span: Exact backtracking and an application," Genes,vol. 12, no. 1, p. 14, Jan2021.

[2] Z. Chen, A. Huang, and X. Qiang, "Improved neural networks based on genetic algorithm for pulse recognition," Computational Biology and Chemistry,vol. 88, p. 107315, Oct 12020.

[3] S. Jain, Y. Tao, and T.Schlick, "Inverse folding with RNA-As-Graphs produces a large pool of candidate sequences with target topologies," Journal of Structural Biology,vol. 209, no. 3, p. 107438, Mar 12020.

[4] T. P.Latchoumi, M. S.Reddy,and K. Balamurugan, "Applied Machine Learning Predictive Analytics to SQL Injection Attack Detection and Prevention," European Journal of Molecular & Clinical Medicine, vol. 7, no. 02, 2020.

[5] B.Tüzün, and E. Saripinar, "Molecular docking and 4D-QSAR model of methanone derivatives by electron conformational-genetic algorithm method,". Journal of the Iranian Chemical Society,vol. 17, no. 5, pp. 985-1000 May2020.

[6] G. Takor, C.E. Morgan, L.Y. Chiu, N. Kendrick, E. Clark, R. Jaiswal, and B.S. Tolbert, "Introducing Structure–Energy Concepts of RNA at the Undergraduate Level: Nearest Neighbor Thermodynamics and NMR Spectroscopy of a GAGA Tetraloop," Journal of Chemical Education,vol. 97, no. 12, pp. 4499-504, Nov 112020.

[7] P.Garikapati, K.Balamurugan, T. P.Latchoumi,and R. Malkapuram, "A Cluster-Profile Comparative Study on Machining AlSi 7/63% of SiC Hybrid Composite Using Agglomerative Hierarchical Clustering and K-Means," Silicon, vol. 13, pp. 961-972, 2021.

[8] T. P.Latchoumi, K.Balamurugan, K.Dinesh,and T. P. Ezhilarasi, "Particle swarm optimization approach for waterjet cavitation peening," Measurement, vol. 141, pp. 184-189, 2019.

[9] A. Khan, H.U. Rehman, U. Habib, and U. Ijaz, "Detecting N6-methyladenosine sites from RNA transcriptomes using random forest," Journal of Computational Science,vol. 47, p. 101238, Nov 12020.

[10] Dhanabalan, S. S., Sitharthan, R., Madurakavi, K., Thirumurugan, A., Rajesh, M., Avaninathan, S. R., & Carrasco, M. F. (2022). Flexible compact system for wearable health monitoring applications.Computers and Electrical Engineering, 102, 108130.

[11] E. Rivas, "RNA structure prediction using positive and negative evolutionary information,"PLoS Computational Biology,vol. 16, no. 10, p. e1008387, Oct 302020.

[12] T. P.Ezhilarasi, G.Dilip, T. P.Latchoumi,and K. Balamurugan, "UIP—A Smart Web Application to Manage Network Environments," In Proceedings of the Third International Conference on Computational Intelligence and Informatics, Springer, Singapore, pp. 97-108, 2020.

[13] J. Cao, and Y. Xue, "Characteristic chemical probing patterns of loop motifs improve prediction accuracy of RNA secondary structures," Nucleic Acids Research,vol. 49, no. 8, pp. 4294-307, May 72021.

[14] M. Salehi, S. Farhadi, A. Moieni, N. Safaie, and H. Ahmadi,"Mathematical modeling of growth and paclitaxel biosynthesis in Corylusavellana cell culture responding to fungal elicitors using multilayer perceptron-genetic algorithm," Frontiers in Plant Science, p. 11, 2020.

[15] R.J. Townshend, S. Eismann, A.M. Watkins, R. Rangan, M. Karelina, R. Das, and R.O.Dror, "Geometric deep learning of RNA structure," Science,vol. 373, no. 6558, p. 1047-51, Aug 272021.

[16] Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.

[17] M.Ramkumar, A. Lakshmi, M.P.Rajasekaran, and A.Manjunathan, "MultiscaleLaplacian graph kernel features combined with tree deep convolutional neural network for the detection of ECG arrhythmia", Biomedical Signal Processing and Control, vol. 76, p. 103639, 2022.