UACEE International Journal of Artificial Intelligence and Neural Networks **ISSN:- 2250-3749 (online)**

RNA Secondary Structure Prediction by Optimization Technique: Genetic Algorithm

GyanPrakashSagar
Department of Computer Science
Punjab Engineering College
Chandigarh, India
gyan.miet@gmail.com

Shailendra Singh
Department of Computer Science
Punjab Engineering College
Chandigarh, India
sscse@in.com

Padmavati
Department of Computer Science
Punjab Engineering College
Chandigarh, India
padma_khandnor@yahoo.co.in

***Abstract*— RNA Secondary Structure prediction is one of the most significant research areas in bioinformatics. This paper presents a RNA secondary structure prediction using the permutation-based Genetic Algorithm (GA). We analyse the selection function STDS and Keep-Best Reproduction (KBR) genetic algorithm replacement techniques, and also using the genetic algorithm crossover operators. We take the SsrS RNA sequences which were the first non-coding RNA to be sequenced, it having the large nucleotides (184nt) that fold into an extended hairpin structure with a large single-stranded internal bulge. We find out the lowest free energy in the individuals RNA sequences with help of the fitness function and which one individual sequence having lowest free energy that individual sequence will be predict the best optimization secondary structure in the RNA individualsequences.**

***Keywords*— Genetic algorithm, RNA secondary structure prediction, Genetic Algorithm representation, and Genetic Algorithm crossover operators.**

I. INTRODUCTION

The optimization problem is the serious issues of this world where we try to find out its solutions by considering various numbers of variables. Genetic information can easily detect through RNA.

RNA is basically stands for ribonucleic acid. The ribonucleic acid (RNA) is an acute molecule in numerous in nature motivating systems. In the biological functions RNA molecule is associated in viral impurity such as the shared cold as well as protein fabrication.In previous few years the RNA secondary structure prediction based is on Nuclear Magnetic Resonance and X-ray crystallography, but these methods are very complicated, time-overwhelming and affluent. So that the mathematical and computational models are developing is very necessary in bioinformatics for RNA secondary structure prediction [1].

In current era genetic algorithm plays a vital role for the development in various fields of science. Basically genetic algorithm is an optimization technique; its basic goal is to find out the best optimum results and solutions.

In this paper we deployed genetic algorithm over various RNA sequences. The various results come out from our experiments which are discussed later in this paper. In second section we explained the permutation-based genetic algorithm involved in our experiment along with that its involvement in our experiment in this section we also explained the RNA secondary structure prediction. In third section we explained the methodology involved in our experiment which explain how we conduct our experiment. In fourth section we explained the final result that comes out from the implementation of genetic algorithm over RNA sequences. Finally we conclude this paper at our last section.

II. RNA SECONDARY STRUCTURE PREDICTION ON PERMUTATION-BASED GENETIC ALGORITHM

Genetic algorithm are search optimization algorithm which is use idea of natural genetics. They start with the random population of individuals solutions. Then each individual is tested for its ability to solve the problem and the process moves to next generation. The use of genetic algorithm to find the set of low energy structures in the RNA molecule [5]. Chromes are the solution of problem in each individual of the population. They use selection, crossover and mutation operators for their evaluation.

A. RNA Secondary Structure Prediction

In the era of a bioinformatics, the ribonucleic acid (RNA) is a naturally kind of molecule. RNA molecule consists of a prolonged chain of nucleotide units. In RNA sequence every nucleotide consists of a nitrogenous base, a ribose sugar and a phosphate [2]. RNA is very similar to DNA but differs a few important structural details: In the biological cell RNA is usually single-stranded, but the DNA is usually double-stranded. RNA nucleotides contain ribose while DNA contains deoxyriboses [3]. An RNA molecule represents a long chain of monomers called nucleotide consist of a base pairs Adenine(A), Cytosine(C), Guanine(G) and Uracil(U).

The structure of RNA is usually modeled as a word over the alphabets A, C, G and U. The two groups of complementary bases A-U and C-G from stable base pairs and are known as the Watson-crick base pair. In this base pair, C-G base pair is more stable than A-U base pair [3, 4]. According to the thermodynamics models that describe the fact that the G-C pair has three hydrogen bonds, the A-U pair has two hydrogen bonds and the wobble pair GU has the weaker bonding than the A-U pair.

B. Permutation Based Genetic Algorithm Representation

In genetic algorithm population type is using as the bit-strings to represent the structure in the population. There are several types of representations and operators of genetic algorithm discussed below.

*1) Individual representation:*In this individual representation is described a secondary structure is an individual in the population and a structure is described as a permutation of helix list [6]. The characteristics of genetic algorithm and they do work with encoded description of an individual, but do not the individual itself.

*2) Fitness Function:*Fitness function describes the negative value of the free energy of a RNA secondary structure. If the secondary structure contains the lowest free energy then that secondary structure has the highest fitness [7].

*3) Initial Generation:*In the initial generation represents a candidate solution when if the set of helices H holds n helices and use a permutation of length n. In the final structure have keeping the order in which a helix present in the permutation is the order in which it is chosen through the decoder. In the structure helices are rejected or removed when the helices are not well-suited with any previously designated helices and the selection probability of helices are proportional to its probability of its computed. If a helix is selected in secondary structure then add it into the permutation representation otherwise reject or remove from the H. After this process we subtract a small positive value B from P_i so that the helix with lower probability has the chance to be selected. When we repeat the step then H is unfilled [6]. The permutation of helix list describes an individual is generated and is also represented. Genetic algorithm is applied after a population of structure is created and applied using the free energy as the measure of fitness until all the structure in the population are constant.

4) Binary Representation: In the genetic algorithm RNA structures can be encoded in the bit-string format. All bit string relates to a helix in the secondary structures. The length of a bit-string is denoted by |H|. The length of bit-string |H| is belonging to the all possible helices in the RNA secondary structures. In the secondary structure, if the helix is present then it is symbolized via a 1 and if the helix is not presented then it is symbolized via a 0. In secondary structure some helices cannot make a suitable structure then using a mechanism is known as the repair mechanism which provides the proper and good structures in the population [8].

5) Permutation: In the permutation RNA structures can be encoded with integer permutation techniques. A permutation technique also having a length |H| like as a binary bit-string. When generating the random permutation for each structure then producing the random population. The process of permutation encoding produced an effective structure by reading the string from left to right. This permutation is basically made by the unique helix identifiers which helix is enclosed in H. The use of this permutation algorithm does not require a repair algorithm saving on computation time. But this algorithm only creates and searches suitable results. This techniques used for any permutation to decode the RNA secondary structure. The using of this permutation encodingtechnique over binary encoding is that permutation crossover operators allow the genetic algorithm to preserve absolute position orders of gens [6].

6) Standard Selection (STDS): It is basically explained by the roulette-wheel selection [103]. In this selection method the sum of the fitness of all the members of the population is associated to each individual is given a pie-shaped slice of a wheel proportional to its fitness. When the wheel stops on its portion then choosing the roulette is rotated and also choosing an individual [13].

7) Keep-Best Reproduction (KBR): It is basically takes two parents and recombines them and the selection method KBR keeps the best parent and also keep the best offspring in order to introduce good new genetic material into the population as well as to keep good old chromosomes. In this KBR operator firstly selects two parents via roulette wheel selection and then apply the crossover and mutation. The rank-based selection passed the best parent and best child to next generation [12, 14, and 15].

C. Genetic Algorithm Crossover Operator

The genetic crossover operators are fairly and easy to implement. Many types of these cross over operators such that scattered, single point, two point, Intermediate, Heuristic and custom etc.

1) Scattered: This operator creates a random binary vector and from the first parent selects the genes where the vector in one and genes where vector is zero from the second parent, and combines the genes to form the child [9].

*2) Single Point:*Asingle point operator selected a casual integer n between 1 and number of variables, and selects the vector entries numbered less than or equal to n from the first parent, and also selects genes numbered greater than n from the second parent, and concatenates these entries to form the child [10].

3) Two Point: Thetwo point crossover method selects two casual integers' m and n among 1 and number of variables. The process picks the genes numbered less than or equal to m from the first parent and chooses genes numbered from m+1 to n from the second parent, and also selects genes numbered

greater than n from the first parent. The process then concatenates these genes to form a single gene [9].

*4) Intermediate:*The intermediate crossover operator generates the children via a casual weighted average of the parents. The intermediate crossover operator is controlled by a single parameter Ratio [10].

III. ALGORITHM

In this genetic algorithm populations of strings are known as chromes, which encode the applicant results known as individuals. The solutions are denoted in binary as string of 0s and 1s but we can use the different type of encoding schemes. This method start from a population of randomly created individuals and it occurs in generation. In this algorithm the fitness value of every individual in the population is calculated. In this method representing the result is an array of bits. The foremost property that makes these genetic representations convenient is that parts are easily aligned due to their fixed size and simple crossover operations.

This genetic algorithm is essentiallystructured by three factors: First is population size, second is crossover probability (P_c) and last is mutation probability (P_m). In this method first we apply the crossover then after mutation is applied on the RNA sequences and this algorithm we download the RNA sequence from the rfam family (http://rfam.janelia.org/browse.html) and this sequence is RF00013 (accession 6s /SsrS RNA). In our experiment we choose the crossover probability P_c = 0.8 and mutation probability is P_m = 0.7. We are using the 25 population sizes, generation count is 900, and 30 numbers of variables per individual are used. In table 1 show the all parameter which we are used in our method. In this algorithm we take the SsrS sequences, the number of individual sequences are take 25 (seq1 to seq25) and we take the 30 number of variables from each 25 SsrSsequences set. We make a fitness function for solve our problem and this solves the lowest minimum free-energy among individual sequences and find the best fitness.

In previous section we already discussed the genetic algorithm and RNA secondary structure.

The experiment that has been conducted are implemented in MATLAB 7.11.0.584(R2010b), Intel core i3, 3 GB RAM, 32 bit window 7 operating system. After the set-up of experiment the result has been shown in next section.

IV. EXPERIMENTAL RESULT

In our experiment we show the best RNA secondary structure prediction among the individuals and basis of minimum free energy with using the fitness function. In permutation-based genetic algorithm we using crossover operator only two point crossover operator. In the figure-1 shows the initial population and initial scores fields in the population panel to the final population of the run before run when the export the problem. The figure shows the best fitness plots from the original run. This best fitness graph is plotted between the generations and the fitness value. In our experiment best fitness is -14.9 and mean fitness is -14.468.

In our experiment objective function value (fval) is -14.39999 and number of function evaluation (funcount) is 1300. On the basis of this objective function is produces minimum free energy (-14.90 kcal/mol). The figure-1 shows the relation between generation and fitness value. This graph shows the best fitness value at -14.9. Our experiment results are show in table-2. The figure-2 shows the selection function between individual and number of children. The figure-3 shows the best optimize secondary structure in the dot bracket notation and figure-4 also shows the best optimize secondary structure in circular form. We can say that all the results are representing the best optimize secondary structure in among individual.

V. CONCLUSION AND FUTURE WORK

In this paper we describe the secondary structure of RNA prediction established on the permutation-based genetic algorithm and using the genetic algorithm crossover operators and this paper representing the best optimizes RNA secondary structure among the allindividual RNA sequences.

The genetic algorithm to optimize the best RNA secondary structure prediction among the individuals RNA sequences on basis of minimum free energy, this minimum free energy is (-14.9000 kcal/mol) and then also having high fitness value. The figure-1 show the best fitness is -14.9 and mean fitness is -14.468. This paper representing the optimization of RNA secondary structure among the RNA individual sequences. This optimization technique shows the best optimize RNA secondary structure prediction on basis of minimum free-energy, among all individual sequences which one have the minimum free energy.

So we can say that, this optimization method genetic algorithm is optimizing the secondary structure in large amount of data sets with the good accuracy, less time and best result.

In future work we will predict the secondary structure with the help of other types of methods like as artificial neural network and support vector machine. We will analyse among these methods and find out which one method is predicting the best secondary structure on basis of taking time and accuracy.

In this genetic algorithm having many type of crossover operators such a single point, two points, scattered, intermediate, heuristic, arithmetic and custom. We will use all different type of crossover operators and comparing among them, and find out the different performance of the crossover operators in RNA sequences.

Table 1 our parameter choices

Parameters	**Parameters values**
Population size	25
Generation Count	900

Number of variables	30
Population Type	Bit-string
Scaling Function	Rank
Replacement	STDS, KBR
Elite count	2
Crossover Function	Two point
Crossover Probability	0.8
Mutation Function	Uniform
Mutation Probability	0.7
Migration Direction	Both
Migration Fraction	0.2

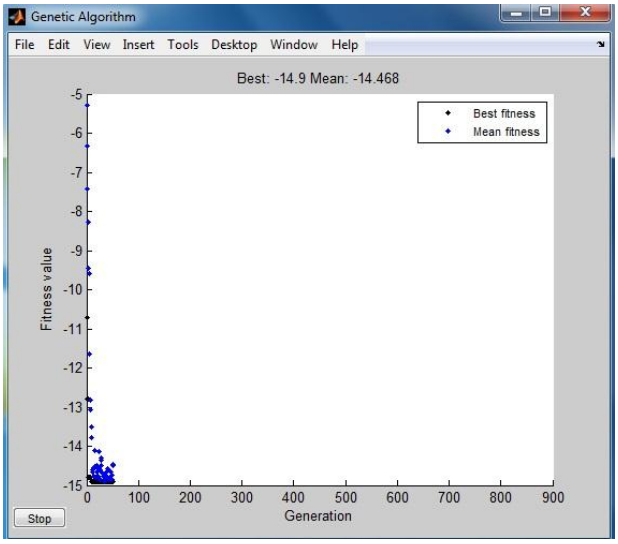

Figure 1shows the best fitness and mean fitness

Table 2 our experimental results

Parameters	**Parameter Values**
Fval	-14.39999
Funcount	1300
Minimum free energy	-14.9000 kcal/mol
Score	Min(-14.90), Max(-9.5000)
Iteration	51
Best-fitness	-14.9
Mean-fitness	-14.468

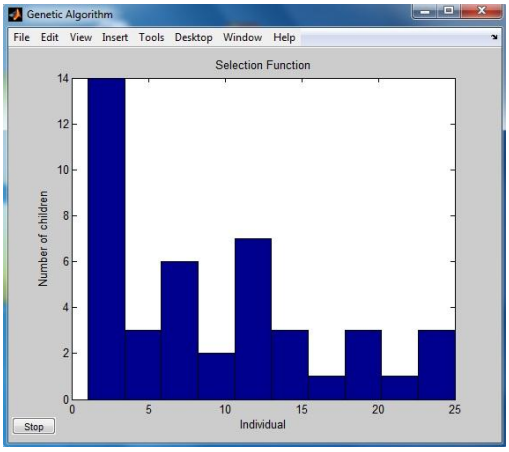

Figure 2 shows the selection function

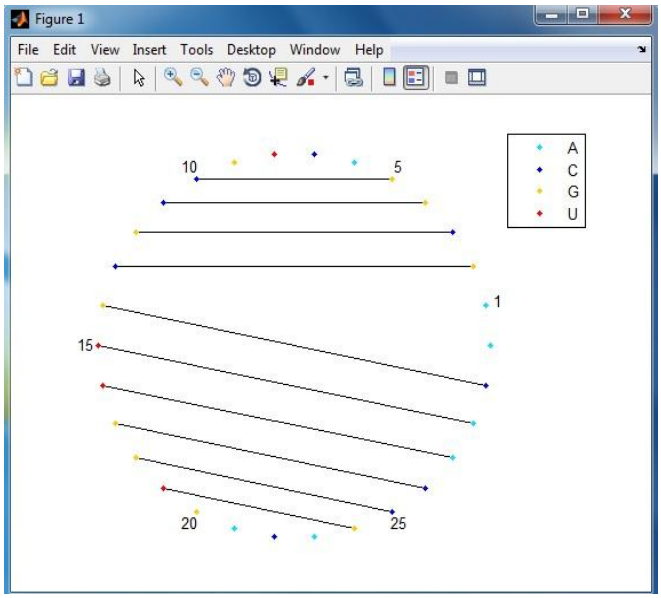

Figure 3 shows the optimize secondary structure in circle

Figure 4 shows the optimize secondary structure in dot bracket

REFERENCES

[1] E. W. Steeg, Artificial Inteligence and Molecular Biology, chapter Neural Networks, Adaptive Optimization, and RNA Secondary Structure Prediction, pp. 121-60, American Association for. Artificial Intelligency, Menlo Park, CA, USA,1993.for. Artificial Intelligency, Menlo Park, CA, USA,1993.

[2] Bohar, H., Bhor, J., Brunak, S., Cotterill, R.M.J., Fredholm, H., Lautrup, B. and Peterson, S.B. (1990) Febs Letters, 261,43-46.

[3] O'Neill, M.C. (1992) Nucleic Acids Res., 20, 3437-3477.

[4] Q. Liu, X. Ye, and Y. Zhang, "A Hopfield neural network based algorithm for rna secondary structure prediction," *Proc. of the First International Multi-Syposiums on Computer and Computational Sciences* (IMSCCS'06), pp. 1-7, 2006

[5] Yuan Xi-min, Li Hong-yan, Li Shu-kun, Cui Guang-tao. The application of Neuran Networks and Genetic Algorithm in water science [m], Bejing. China Water Conservancy and Hydropower Press.2002,8.

[6] K. C. Wiese, E. Glen. "A permutation Based Genetic Algorithm for the RNA Folding Problem: A Critical Look at Selection Strategies, Crossover Operators and Representation Issues", BioSysrem-Special Issue on Computational intelligence in Bioinformatics, Fogel G, Corne d, (eds.) in press, 2003.

[7] Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H.: Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Bio., 288:911-11940, 1999.

[8] F. H. D. van Batenburg. A. P. Gultyaev, and C. W. A. Pleij, "An APL-programmed genetic algorithm for the prediction of RNA secondary structure," *Jouranal of Theoretical Biplogy*, vol. 174,pp. 26-280, 1995.

[9] D. E, Goldberg, R. Lingle. "Alleles, Loci, and the Travelling Salesman Problem," in Proceedings of the International Conference on Genetic Algorithms and Applications, 1985.

[10] I. M. Oliver, D.J. Smith, and J. R. C. Holland, "A Study of Permutation Crossover Operation On The Travelling Salesman Problem," in *Proceeding of the Second International Conference on Genetic Algorithms*, 1987.

[11] T. Starkweather, S. McDanial, C. Whitely, K. Matheas, and D. Whitely, "A comparison of genetic sequencing operators," in *Proceeding of the Fourth International Conference on Genetic Algorithms*, R. Belew and L. Booker, Eds, Los Altos: Morgan Kaufmann Publishers, 1991, pp. 69-76.

[12] K. C. Wiese and S. D. Goodwin, "Keep-Best Reproduction: A Local Family Competition Selection Strategy and the Environment it Flourishes in." *Constraints*, vol. pp. 399-422, 2001.

[13] J. E. Baker. Reducing bias and inefficiency in the selection algorithm. In J. J. Grefenstette, editior, *Proceeding of the Second International Conference on Genetic Algorithms and their Application*, pages 14-21, Hillsdale, New Jersey, USA, 1987. Lawrence Erlbaum Associates.

[14] Kay Wiese and Scott D. Goodwin. Convergence characteristics of keep-best reproduction. In SAC '99. *Proceeding of the 1999 ACM Symposium on Applied Computing 1999*, pages 312-318. ACM, 1999.

[15] Kay Wiese and Scott D. Goodwin. Keep-best reproduction: A local family competition selection strategy and the environment it flourishes in. *Constraints*, 6(4):399-422, 2001