Genetic Algorithm based Optimizer for RNA Secondary Structure Prediction

Gyan Prakash Sagar
Department of Computer Science
Punjab Engineering College
Chandigarh, India

Shailendra Singh
Department of Computer Science
Punjab Engineering College
Chandigarh, India

Padmavati
Department of Computer Science
Punjab Engineering College
Chandigarh, India

ABSTRACT
The paper represents the optimization of RNA secondary structure prediction in RNA molecule by using the genetic algorithm based on binary crossover operators. Using the selection function STDS keep best reproduction for RNA prediction. Take the number of individual RNA sequence sets from the Rfam family database and calculate the lowest free-energy in the individual RNA sequence sets. Apply the fitness function for optimize the lowest free-energy in the each individual sequence sets. Which one individual sequence set have the lowest free-energy that individual sequence set will be predict the best optimizer secondary structure in the RNA individual sequence and used the RNA fold algorithm for calculating the free-energy in the RNA molecules.

Keywords
Genetic Algorithm, RNA secondary structure, RNA folding, minimum free-energy, Genetic algorithm representation, Genetic algorithm crossover operators.

1. INTRODUCTION
The role of RNA is suffered a foremost in the past few years. The most folding algorithm like as dynamic programming methodology the unusual of RNA structure prediction as a global optimization problem. The goal of the global optimization to find out the lowest free-energy from the RNA secondary structure and which one structure having lowest minimum free-energy producing best stable secondary structure among them. In the biological cell, the biomolecules are not in a conformation that has the lowest energy and they occupy a number of conformations that are at thermal equilibrium with each other.

In bioinformatics field the current epoch of genetic algorithm plays a vigorous role for the enlargement and improvement in prediction of RNA secondary structure. Generally the genetic algorithm is an optimization technique and its simple goal is to find out the best optimum results in given problems. In biological sequences the improved genetic algorithm assessed more true positive stem loops and base pairs than viable with dynamic programming algorithm [1].

In this paper we apply the genetic algorithm on the large amount of RNA sequences. In our experiment the various results find out and these results are discussed later in this paper. In second section we explained the 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 demeanour 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 in conclusion.

2. RNA SECONDARY STRUCTURE PREDICTION BASED ON GENETIC ALGORITHM
Genetic algorithm is based on the search optimization algorithm in era of bioinformatics which is use idea of natural genetics and the populations of individual's solutions are randomly started. The genetic algorithm provides the two types of characteristics maximization and minimization. The search optimization algorithm each individual is verified for its knack to solve the problem and the process passages to next generation. The use of genetic algorithm in the era of bioinformatics 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. In this genetic algorithm we use the selection function, crossover operators and mutation function for their assessment.

2.1 RNA Secondary Structure prediction
In the biological sequences, the ribonucleic acid (RNA) is a naturally generous of molecule. RNA molecule consists of an overlong chain of nucleotide units. In RNA sequences every nucleotide consists of a nitrogenous base, a ribose sugar and a phosphate [2]. RNA is very similar to DNA in biological sequences but different of between some significant structural specifics. The structure of biological cell, the RNA is ordinarily single-stranded and the DNA is ordinarily double-stranded. In biological molecule RNA nucleotides contain ribose however DNA contains deoxyiboses [3]. An RNA molecule represents a long chain of monomers called nucleotide and the building blocks of RNA are four types of nucleotides: Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). The scaffold of the tertiary structure is determined by the secondary structure.

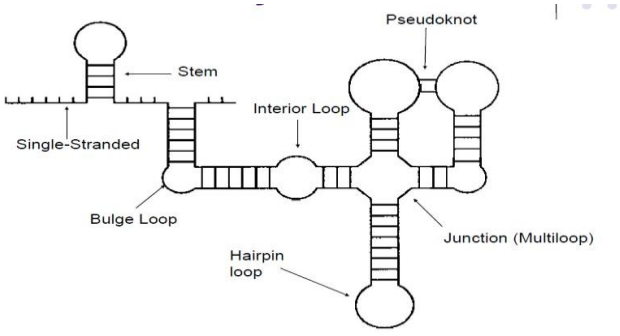

Fig 1 RNA stacking region and loop elements

The structure of RNA is usually described as a word over the alphabets A, C, G and U. In RNA structure A-U and C-G are the two groups of complementary bases from stable base pairs and these stable structure are known as the Watson-crick base pair. In this stable structure the base pair, C-G is more stable than A-U base pair [3, 4]. According to the thermodynamics prototypes that explained the fact the three hydrogen bonds in the G-C base pair, two hydrogen bonds in the A-U base pair and the wobble pair G-U has the weaker bonding than the A-U pair in the RNA molecule. In figure-1 the most common loop motifs are hairpin loops, bulges and internal loops. The RNA secondary structure finding from the RNA sequences can be classified as a global optimization problem and there are many different algorithms that try to solve it starting from the RNA sequence [9, 10].

2.2 Representation of Genetic Algorithm

The genetic algorithms are one of the best ways to solve an optimization problem for which miniature is recognized. In any search space these algorithms can do work. A genetic algorithm will be able to generate an extraordinary superiority justification and use the principles of selection and evolution to produce some clarifications to a given problem. The genetic algorithm containing features of like as individual representation, fitness function, initial generation, permutation, standard selection (STDS) and keep-best reproduction (KBR) [12, 13,14 and 15].

2.3 Genetic Algorithm Crossover Operators

The genetic crossover operators are legitimately and stress-free to implement. In genetic algorithm many types of the crossover operators like as scattered, single point, two point, intermediate, heuristic and custom etc. In single point operator chooses a casual integer n between 1 and number of variables, and then selects the vector items numbered less than or equal to n from the first parent, selects genes numbered greater than n from the second parent, and concatenates these entries to form the child [10].The two point crossover method selects two spontaneous integers' m and n among 1 among number of variables. The method preferences the genes numbered less than or equal to m from the first parent and takes genes numbered from m+1 to n from the second parent, and also takes the genes numbered greater than n from the first parent. The method then concatenates these genes to form a single gene [9].

3. PROPOSED ALGORITHM

The genetic algorithm is produces an array of random numbers that is known as the initial population. In this generated array, each row is signifying of one set of values for the variables. All rows are goes to as individuals in this method. All variable is denoted in this method, the all individual by a set of genes and these genes set is known as a chromosome. These chromosome having many genes but we can say that each variable is ultimately only one number. In a chromosome when all genes are combined into a single value known as phenotype. Therefore the initial population is prompted and then after each individual must be compact to its phenotype form and then after this process this form is served through a fitness function which assigns a fitness value to each individual based on characteristics of its own chromosomes. A fitness function is in genetic algorithm, it is a type of objective function and to propose the optimality of a solution.

In this genetic algorithm have mainly three parameters: First is population size, second is crossover probability (P_c) and last is mutation probability (P_m). The proposed algorithm used the minimization property of genetic algorithm. The proposed method takes RNA sequence from Rfam family database into the fasta format and the sequence accession number is R00002 (5.8S ribosomal RNA) and then encoded into integer format (A=1, C=2, G=3, U=4). First apply the binary crossover operators then after mutation is applied on the RNA sequences. The used of fitness function in proposed method for decoding, integer to RNA alphabet (1=A, 2=C, G=3, G=4) and calculate the free-energy. Which sequence set has minimum free-energy that predicts the best optimize RNA secondary structure among the all RNA sequence sets. The following given the sequence is a type of sequence set. The sequence sets are taking (seq1 to seq20) and seq1 shows the example of sequence set which used in proposed algorithm.

```
seq1='GGGGAUGAUCGGUUUCGACAUUGCCUGCAAAA
CUGUGAGAAGCGGGUCGAGAAUGCAGCCUUAUCUCG
UUAACGAUGACUGCAAACUAUAAGUGCCAAUUCAAA
GCGCACUGACUUCGCCCUCGCUGCCUAAGCGAGCGCA
CUAAAGAAGUCCGUCAGACCGGGAAUGCUCUCUACC
CGGAUCCUGGCGCAAUUUAGAGAGAUUGCUGCGUAG
UUACGCCUGAGGGCUACGCGGGACUUGCACUCUGGC
UGGGCUUGUUGAUCUAGGUGCUUGUGGCAAAGAUCA
GAGCCGAGUAGAACGUCUGCACAAGCUACACCCGUA
GAAGGCACAGAAUUACAGCAGUGGACGGGGGUUCAA
UUCCCCCCAUCUCCACC';
```

In proposed algorithm choose the crossover probability $P_c = 0.9$, mutation probability is $P_m = 0.8$ and elite count is 1 and one point crossover operator. The work for predicting the secondary structure, using the 20 population sizes; generation count is 700, and taking 20 numbers of variables per individual sequences. The proposed method takes the sequences (seq1…..seq20) represent the set of variables and also take the 20 number of variables per individual sequence sets or variables. In the algorithm used the RNA fold method for calculating the free-energy in the RNA molecules. In table 1 show the all parameter which are used in the method. After these operations choose the minimum free-energy in the RNA molecules, and which one RNA molecule has a lowest minimum free-energy that will optimize the best secondary structure with having the highest fitness.

In previous section 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 this operation of research the result has been shown in the next section.

4. EXPERIMENTAL RESULT

In the experimental work show that the best optimized RNA secondary structure prediction among the individual RNA sequences or sequence sets. This optimized result prediction on the basis of minimum free-energy in RNA molecules. Table-2 shows the experimental results.

In this experiment, objective function value (fval) is -9.900 and number of function evaluation (funcount) is 1060, minimum score is (-9.9000) and maximum score is -6.2000. On the basis

of this objective function is produces minimum free energy (-9.900 kcal/mol). The Figure-2 shows the best optimization RNA secondary structure with minimum free energy (-9.9000 kcal/mol) among the all individual or among the 20 population size and 200 generation count. The Figure-2 also describe the best optimize RNA secondary structure in the form of dot bracket notation among all individual RNA sequence sets. The Figure-3 shows the best fitness (-9.9), which has minimum free-energy and mean fitness (-9.1) among the 200 generations and also represents the fitness value of 200 generation in RNA molecule. This figure shows the best optimizer for RNA secondary structure prediction among the generations and fitness values. Figure 4 describe the average distance 0f 200 generation which produce the average distance between individuals of RNA sequences. It is used for calculating distance of each individual. Figure 5 describe the current best individual of 20 variables of RNA molecules and each variable represent the personal best fitness value. Figure 6 described the score diversity histogram of number of 20 number of individuals which represent between -9 to -8 on the x-axis. .

5. CONCLUSION AND FUTURE WORK

In this work describe the overview of RNA secondary structure prediction based on the genetic algorithm and using the various genetic algorithm crossover operators. In this experiment find out the best optimize RNA secondary structure among all individual RNA sequence sets.

The genetic algorithm to optimize the best RNA secondary structure prediction among the individual RNA sequences on basis of minimum free energy, this minimum free energy is (-9.7000 kcal/mol) and the optimize RNA secondary structure is shown in the figure-2 in the dot bracket notation. Among all individuals sequences only any one individual sequence set which having the lowest minimum free energy that individual sequence having a highest-fitness and that optimize the best RNA secondary structure. In figure-3 shows the best fitness between generations and fitness values. So we find the best optimize results of our problems with the help of genetic algorithm and the help of its representations and crossover operators.

In future work using the all type of crossover operators. The others crossover operators like as Mapped Crossover (PMX), Order Crossover (OX), and Cycle Crossover (CX) are applied on the RNA sequences and find out the minimum free energy per an individual sequence. At last we compare the all type of crossover methods and find out the, which one crossover techniques will give the best optimum results for RNA secondary structure prediction.

Table 1 Parameter Choices

Parameters	Parameters values
Population size	25
Generation Count	200
Number of variables	20
Creation Function	Uniform
Initial Range	[0; 5]
Population Type	Bit-string
Fitness Scaling Function	Top
Fitness Quantity	0.6
Replacement	STDS, KBR
Elite count	1
Crossover Function	One point
Crossover Probability	0.9
Mutation Function	Uniform
Mutation Probability	0.8
Mutation Rate	0.04
Migration Direction	Forward
Migration Fraction	0.2
Migration Interval	40

Table 2 Experimental Results

Parameters	Parameter Values
Fval	-9.9
Funcount	1060
Minimum free energy	-9.7000 kcal/mol
Score	Min(-9.900), Max(-6.2000)
Iteration	100
Best-fitness	-9.9
Mean-fitness	-9.1
Exit flag	1

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

Figure 2 Optimize RNA Secondary Structure

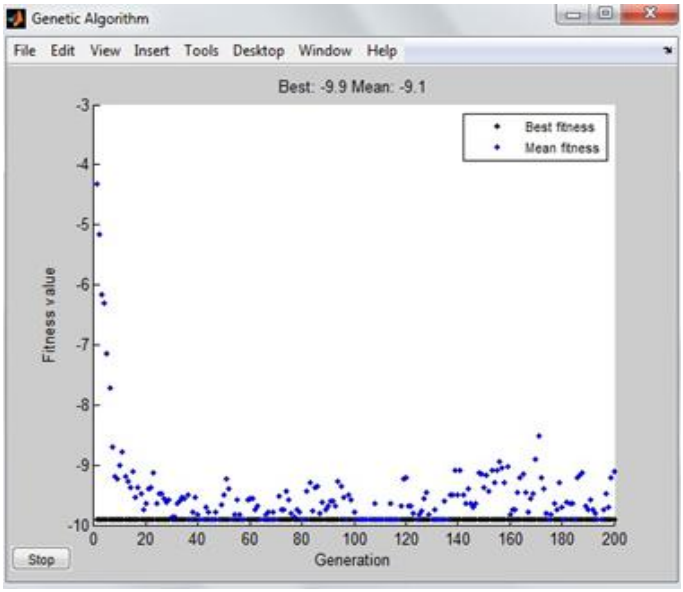

Figure 3 Fitness value of each generation

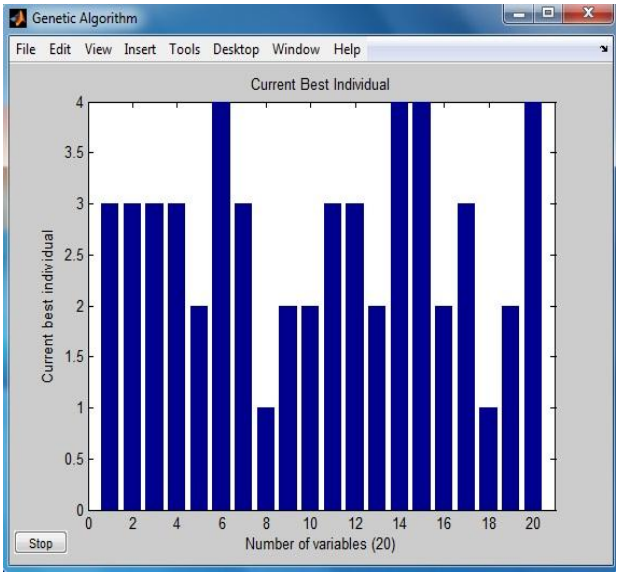

Figure 5 Fitness value of each variable

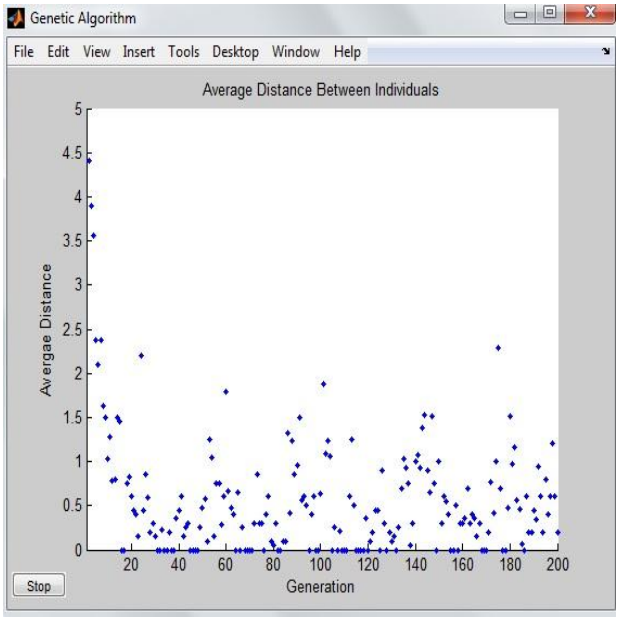

Figure 4 Average distance of each Generation

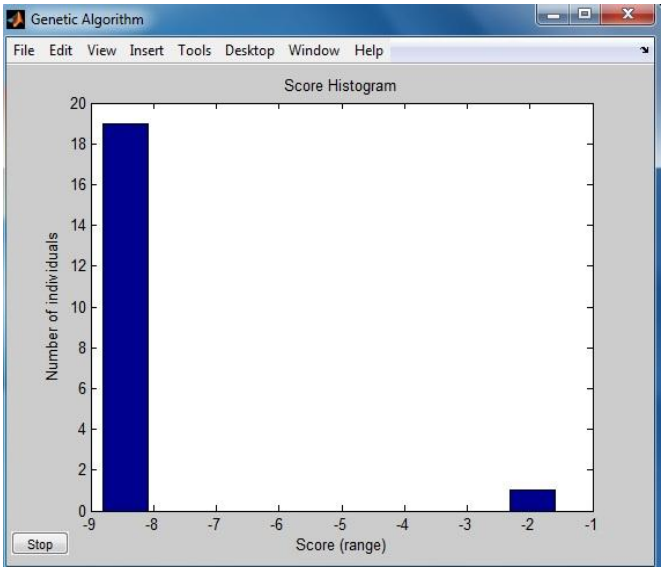

Figure 6 Diversity histogram of each individual

6. REFERENCE

[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] M. Zukar. Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction. *Nuc. Acid. Res.*, 31:3406, 2003.

[10] B. A. Shapiro and J. Navetta. A Massively Parallel Genetic Algorithm for RNA Secondary Structure Prediction. *J. supercomput*, 8:195-207, 1994.

[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