New Approach in Genetic Algorithm for RNA Secondary Structure Prediction

Binh Doan Duy
Faculty of Information Technology, Danang University of Science and Education, the University of Danang, Vietnam
Email: ddbinh@ued.udn.vn

Minh Tuan Pham
IT Faculty, Danang University of Science and Technology, the University of Danang, Vietnam
Email: pmtuan@dut.udn.vn

Long Dang Duc
VN-UK Institute for Research & Executive Education, the University of Danang, Vietnam
Email: long.dang@vnuk.edu.vn

Hoan Dau Manh
Learning and Research Center, Quang Binh University, Vietnam
Email: daumanhhoan@yahoo.com

***Abstract*—RNA secondary structure problem is one of the most important fields in computational molecular biology. Ribonucleic Acid (RNA) has important structural and functional roles in the cell and plays roles in many stages of protein synthesis. The structure of RNA largely determines its function. Many methods can be used to predict the secondary structure of an RNA molecule. One of the methods is the dynamic programming approach. However, the dynamic programming approach usually takes too much time. Genetic algorithm is an evolutionary approach for solving space layout and optimization problems. Due to some drawbacks in genetic algorithm, several modifications are performed on this algorithm. When the advantages of GA are combined with advantages of another algorithm then this approach is called Hybrid Genetic Algorithm. In this paper we introduce a hybrid Genetic Algorithm with Fuzzy Logic. From this hybrid algorithm we apply the problem of predicting the secondary structure of Ribonucleic Acid. Problem predicted secondary structure that we mentioned methods based on thermodynamics, to find secondary structure with minimum energy. With the results of the algorithms found by the hybrid algorithm we introduce, we hope to contribute to the molecular biology data warehouse for molecular biology research. It also introduces a new approach to genetic algorithms.**

***Index Terms*—genetic algorithm, dynamic programming algorithms, minimum free energy, fuzzy logic, RNA secondary structure, computational biology**

I. INTRODUCTION

RNA is nucleic acid consist of a long linear polymer of nucleotide units found in the nucleus. RNA is similar to DNA but usually consists of a single strand instead of double stranded, containing ribose rather than deoxyribose, and has the base Uracil (U) in place of Thymine (T). There are a various form of RNA are found: heterogeneous nuclear RNA (hnRNA), messenger RNA (mRNA), transfer RNA tRNA), ribosomal RNA (rRNA), and small nuclear RNA. Structurally, hnRNA and mRNA are both single stranded, while rRNA and tRNA form three-dimensional molecular configurations. Each type of RNA has a different role in various cellular processes such as carrying genetic information (mRNA), interpreting the code (rRNA), and transferring genetic code (tRNA). It also performs different functions which include: catalyzing chemical reactions [1], [2], directing the site specific modification of RNA nucleotides, controlling gene expression, modulating protein expression and serving in protein localization [3].

RNA is a single strand of nucleotides composed of Adenine (A), Guanine (G), Cytosine (C) and Uracil (U) and it can fold back on itself to form its secondary structure with base pairs like A ≡ U, G = C, and G - U. However, an RNA sequence can fold to form several possible secondary structures. Determining which is the correct secondary structure is called the *RNA secondary structure prediction problem* [4].

The function of RNA molecules determines many diseases caused by RNA viruses. Identifying the secondary structure of an RNA molecule is the fundamental key to understand its biological function [5], [6].

The structure of an RNA molecule can be crucial for its function. Accordingly, the automatic prediction of RNA structures from sequence information is an important problem. Today, there are two prediction strategies:

Manuscript received April 22, 2020; revised September 23, 2020.

doi: 10.12720/jait.11.4.249-258

- **Thermodynamic approaches:** The conformation of paired and unpaired regions in an RNA structure can be associated with an energy value. Given some energy model, thermodynamic approaches find the energetically most stable structures among all possible secondary structures of an RNA sequence. Such a structure is denoted the Minimum Free Energy (MFE) structure.
- **Comparative approaches:** In functional noncoding RNA, the structure of an RNA is conserved during evolution. Since a base pair can be formed by different combinations of nucleotides, different sequences can have the same or a similar structure. If a family of structural homolog RNA molecules has a sufficient amount of sequence conservation, a multiple sequence alignment can emphasize regions of sequence variation. The regions containing structure-neutral mutations, denoted as compensatory base changes, give clues to the structure of an RNA molecule.

A common computational approach is to find the secondary structure with the Minimum Free Energy (MFE), relative to the unfolded state of the molecule. There is considerable evidence that RNA secondary structures usually adopt their MFE configurations in their natural environments [7].

The forces driving RNA folding can be approximated by means of an energy model, which contains a set of model features, corresponding to small RNA structural motifs, and model parameters. Each parameter associates a free energy change value with a model feature. Current energy-driven computational approaches take as input an RNA sequence, and find a structure which optimizes an energy function, using a given energy model.

For secondary structure prediction, a genetic algorithm is more practical since it can solve a problem of large size. In this paper, we shall solve the problem of RNA secondary structure prediction with genetic algorithms combined with fuzzy logic. The organization of this paper is as follows. In Section II, we shall introduce the RNA secondary structure and prediction. In Section III, we shall present Genetic algorithm. We describe the Fuzzy logic in Section IV. The proposed hybrid GA-Fuzzy logic technique is described in Section V. The results and discussions are presented in Section VI. Conclusion remarks are done in Section VII.

II. RNA Secondary Structure and Prediction

A. RNA Secondary Structure

RNA molecules are characterized by sequences of four types of nucleotides or bases: Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). The linear base sequence of an RNA strand constitutes the primary structure or sequence, and is formally defined as follows:

Definition 1. An RNA sequence of length n nucleotides is a sequence $x = x_1 x_2 \ldots x_n$, where $x_i \in A, C, G, U, \forall i \in 1, \ldots, n$.

An RNA sequence tends to fold to itself and form pairs of bases. The set of base pairs that form when an RNA sequence folds is called RNA secondary structure, defined as follows:

Definition 2. An RNA secondary structure y compatible with an RNA sequence x of length n is defined as a set of (unordered) pairs (s,t) , with $s,t \in 1,\ldots n$ that are pairwise disjoint, i.e., for any two pairs (s,t) and $(u,v) \in y, (s,t) \cap (u,v) = \varnothing$ (the empty set).

Thus, in an RNA secondary structure, each base can be either unpaired or paired with exactly one other base. The base pairs of a secondary structure arise mainly because of the stability of the hydrogen-bonding between bases, stacking interactions with adjacent nucleotides, and entropic contributions. The most common hydrogen bonds which lead to secondary structure formation are between C and G, between A and U (both pair types are called Watson-Crick pairs), and between G and U (called wobble pairs). The stability of these base pairs is given by the following relation: C-G > A-U ≥ G-U [8], [9]. Throughout this thesis, we consider that all C-G; A-U and G-U base pairs are canonical, and all other base pairs are non-canonical. However, we note that from the point of view of the planar edge-to-edge hydrogen bonding interaction [8], there are C-G, A-U and G-U base pairs that do not interact via Watson-Crick edges, and there are non-canonical base pairs that do interact via Watson-Crick edges [8].

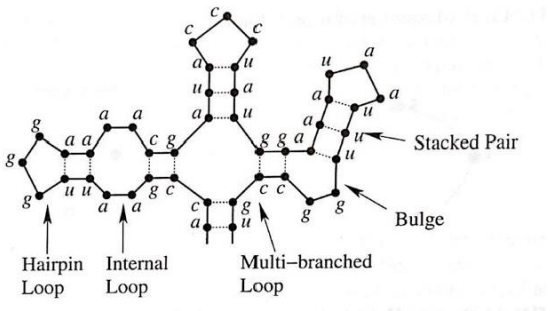

Figure 1. A pseudoknot-free secondary structure.

The structural motifs we consider in this work are the following:

- A *stacked pair* contains two adjacent base pairs. A stem or helix is made of one or more adjacent base pairs. The stem marked in Fig. 1 has one stacked pair or two base pairs.
- A *hairpin loop* contains one closing base pair, and all the bases between the paired bases are unpaired.
- An *internal loop*, or interior loop, is a loop having two closing base pairs, and all bases between them are unpaired.
- A *bulge loop*, or simply bulge, is a special case of an internal loop that has no free base on one side, and at least one free base on the other side.
- A *multibranch loop*, multi-loop, or junction, is a loop that has at least three closing base pairs;

stems emanating from these base pairs are called multi-loop branches.

- The *exterior loop*, or external loop, is the loop that contains all the unpaired bases that are not part of any other loop.
- The *free bases* immediately adjacent to paired bases, such as in multi-loops or exterior loops, are called dangling ends.

If a secondary structure contains only the aforementioned motifs, it is called pseudoknot-free. A formal definition follows:

Definition 3. A pseudoknot-free RNA secondary structure y compatible with an RNA sequence x of length n is an RNA secondary structure in which any two pairs (s,t) and $(u,v) \in \mathrm{y}$ are either nested, i.e., $s < u < v < t$, or follow each other, i.e. $s < t < u < v$. Here we have assumed without loss of generality that $s < t, u < v$ and $s < u$.

A pseudoknot is a structural motif that involves non-nested (or crossing) base pairs (see details below). Fig. 2 contains one pseudoknot, and the structure is called pseudoknotted secondary structure, with the following definition:

Definition 4. A pseudoknotted RNA secondary structure y compatible with an RNA sequence x of length n is an RNA secondary structure in which there exist at least two base pairs (s,t) and $(u,v) \in \mathrm{y}$, for which $s < u < t < v$ (these are often called crossing base pairs). Here we have assumed without loss of generality that $s < t, u < v$ and $s < u$.

Figure 2. RNA secondary structure with simple pseudoknots.

B. RNA Secondary Structure Prediction

The problem of RNA secondary structure prediction can be formalized as follows:

- Given: an RNA sequence x and a free energy model M;
- Objective: develop an algorithm $A(x,M)$ that returns one or more RNA secondary structures y compatible with x that are predicted to be of biological interest.

A common approach to obtain biologically interesting secondary structures (i.e., native or functional secondary structures) is to find the minimum free energy (MFE) configuration y^{MFE} of a given RNA sequence x under the assumed free energy model M. This approach is based on the assumption that RNA molecules tend to fold into their minimum free energy configurations,

$$y^{MFE} \in \arg\min_{y \in Y} \Delta G(x, y, M) \quad (1)$$

where Y denotes the set of all possible pseudoknot-free secondary structures for x, ΔG is an energy function that gives a measure of folding stability, and argminy $\Delta G(y)$ denotes the (set of) y for which $\Delta G(y)$ is minimum.

Since a pseudoknot-free secondary structure can be decomposed into several disjoint pseudoknot-free structures with additive free energy contributions, dynamic programming algorithms are suitable for this problem. The dynamic programming algorithm of Zuker and Stiegler [10] starts from hairpin loops, and recursively fills several dynamic programming arrays with the optimal configuration for subsequences delimited by every possible base pair (s,t), where $1 \le s,t \le n$ and n is the length of x. This algorithm is guaranteed to find the minimum free energy pseudoknot-free secondary structure for a given RNA sequence in $\Theta(n^4)$ (or $\Theta(n^3)$) if the number of unpaired bases in internal loops is bounded above by a constant, or if the later extension of Lyngso *et al*. This algorithm and various extensions of it are implemented in a number of widely used software packages such as Mfold [10], RNA structure [11], the Vienna RNA Package [12].

C. RNA Thermodynamics and Free Energy Models

1) RNA thermodynamics: The stability of an RNA secondary structure is quantified by the free energy change ΔG, measured in kcal/mol. The free energy G indicates the direction of a spontaneous change, and was introduced by J. W. Gibbs in 1878 [13]. The free energy change ΔG quantifies the difference in free energy between the folded state of the molecule and the unfolded state. ΔG represents the work done by a system at constant temperature and pressure when undergoing a reversible process. A folded RNA has negative free energy change, and the lower it is, the more stable the structure is. The base pairs are usually favorable to stability, while the loops are usually destabilizing. The free energy change is a function of enthalpy change ΔH, entropy change ΔS and temperature T (in Kelvin), according to the Gibbs function:

$$\Delta G = \Delta H - T.\Delta S \quad (2)$$

Enthalpy (H) is a measure of the heat flow that occurs in a process. The enthalpy change (ΔH) for an exothermic reaction, such as RNA folding, (i.e., the heat flows from the system to the surroundings) is negative. The enthalpy is measured in kcal/mol. The formation of RNA stems is the dominant enthalpic factor, through hydrogen bonding and stacking interactions.

Entropy (S) is widely accepted as a thermodynamic function which measures the disorder of a system. Thus, the entropy change ΔS measures the change in the degree of disorder. If ΔS is positive, it means there was an increase in the level of disorder. A negative value indicates a decrease in disorder.

However, a modern view of the entropy change presents it as the quantity of dispersal of energy per temperature, or by the change in the number of microstates: how much energy is spread out in a process, or how widely spread out it becomes - at a specific temperature. If ΔS is negative, such as for RNA loops, it means the amount of energy dispersed decreased. The loops in an RNA structure contribute to the entropy more than to the enthalpy because the folding process restricts the microstates of the loop nucleotides as compared to the unfolded strand. The entropy is measured in $kcal/(mol\ K)$ or entropy units ($1eu = 1cal/(mol\ K)$).

2) RNA free energy models: An RNA free energy model is a theoretical construct that represents the rules and variables according to which RNA sequences form (secondary) structures. We consider an RNA free energy model that has three main components:

1) A collection of structural features $(f_1, f_2, ..., f_p)$, where p is the number of features of the model. A feature is an RNA secondary structure fragment whose thermodynamics are considered to be important for RNA folding.
2) Collection of free energy parameters $(\theta_1, \theta_2, ..., \theta_p)$, with free energy parameter θ_i corresponding to feature f_i. The parameter θ_i is sometimes denoted by $\Delta G(f_i)$.
3) A free energy function that defines the thermodynamic stability of a sequence x folded into a specific secondary structure y that is consistent with x. Most models for pseudoknot-free secondary structure prediction assume that the free energy function of sequence x and structure y is linear in the parameters θ_i, of the form:

$$\Delta G(x, y, \Theta) := \sum_{i=1}^{p} c_i(x, y).\theta_i = c(x, y)_{\mathrm{T}}\Theta \quad (3)$$

where $\Theta := (\theta_1, \theta_2, ..., \theta_p)$ denotes the vector of parameter values $\theta_i, c_i(x, y)$ is the number of times feature f_i occurs in secondary structure y of sequence x, and $c(x, y) := (c_1(x, y), ..., c_p(x, y))$ denotes the vector of feature counts $c_i(x, y)$

The main feature categories:

- Stacked pair features $stack(a, b, c, d)$ where $a-d, b-c$ form base pairs. For example, the feature marked 1 in Fig. 3 corresponds to a=A; b=G; c=C; d=U. ThecTurner99 [14], [15] free energy value for this feature is -2.1 kcal/mol.
- Hairpin loop terminal mismatch features $HLtm(a, b, c, d)$ where $a-d$ forms the hairpin loop closing base pair, and b and c are the first two unpaired bases of the hairpin loop, forming stacking interactions with the closing base pair. For example, in the hairpin loop marked by 2 in Fig. 3, a=C; b=G; c=A and d=G. The Turner99 value for this feature is -2.2kcal/mol.
- Features for 1×1; 1×2 and 2×2 internal loops, where 1×1 means there is one unpaired nucleotide on each side of the internal loop, and 1×2 and 2×2 have analogous interpretations. Experiments show that often the unpaired bases in a small internal loop form hydrogen bonds and other interactions, and these bases are sometimes considered to form noncanonical base pairs. The thermodynamics of such internal loops do not obey the nearest-neighbour principle [15], therefore the Turner model includes sequence dependent features for them.
  - For 1×1 internal loops, the features are $IL11(a, b, c, d, e, f)$, where $a-f$ and $c-d$ form base pairs, and b and e are the unpaired (also called noncanonically paired or $b-e$ mismatch). The internal loop marked with 3 in Fig. 3 falls into this category, where a=G; b=A; c =C; d=G; e=G; f=C. The Turner99 value is 0.4 kcal/mol.
  - For 1×2 internal loops, the features are $IL12(a, b, c, d, e, f, g)$, see for example the internal loops marked with 4 in Fig. 3.
  - For 2×2 internal loops, the features are $IL22(a, b, c, d, e, f, g, \mathrm{h})$, see for example the internal loop marked with 5 in Fig. 3.
- Internal loop terminal mismatch features $ILtm(a, b, c, d)$, where $a-d$ is one of the closing base pairs for a general internal loop, and b and c are the first unpaired nucleotides adjacent to the closing base pair. For example, in the internal loop marked by 6 in Fig. 3, there are two terminal mismatches, each one corresponding to each closing base pair and the adjacent unpaired nucleotides. For the leftmost one a=C; b=G; c=A and d=G, with Turner99 value of -1.1 kcal/mol.
- Features for the number of unpaired nucleotides in hairpin loops, internal loops and bulge loops: $HLlength(l), ILlength(l)$ and $BLlength(l)$ ur where l is the number of unpaired nucleotides in the loop. For example the hairpin loop marked 2 in Fig. 3 has length 5 (with Turner99 value 1.8 kcal/mol), and the internal loop marked 6 has length 7 (with Turner99 value 2.2 kcal/mol).
- Three multi-loop features: $Multi-a$ is the multi-loop initiation, $Multi-b$ is the multi-loop number of branches, and $Multi-c$ corresponds to the number of unpaired bases in a multi-loop. The Turner99 parameter values for these three

features are 3.4, 0.4 and 0.0 kcal/mol, respectively. For example the multi-loop marked 7 in Fig. 3 has three branches and six unpaired bases. In the Turner99 model, (in addition to other terms) this multi-loop contributes a linear energy function of these three parameters to the total free energy function: $1.Multi-a+3.Multi\_b+6.Multi-c$.

- Dangling end features: $dangle5(a,b,c)$ where $b-c$ forms a base pair, and a is a base towards to 5 end of the molecule. For the dangling end marked 8 in Fig. 3, a=C; b=G and c=C, and the Turner99 value is -0.3 kcal/mol. Similarly, dangle3(a; b; c) are features for an unpaired base c adjacent to a base pair $a-b$, towards the 3 end of the molecule. For the dangling end marked by 9, a=C; b=G and c=U, and the Turner99 value is -0.6 kcal/mol. In the Turner99 model, the dangling end features are included in the energy contribution of multiloops and exterior loops.
- Other features, including special cases of stable hairpin loops, asymmetric internal loops and penalty for intermolecular initiation for the case of interacting RNA molecules.

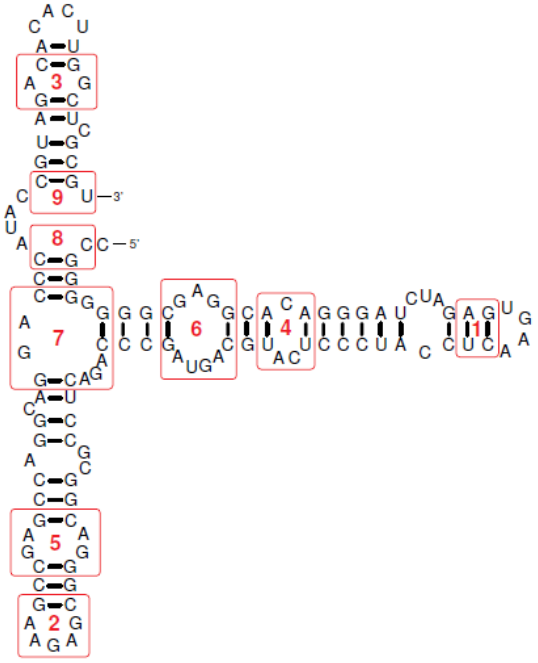

Figure 3. Secondary structure of an arbitrary RNA sequence. Marked in red boxes are RNA structural motifs, including stacked pairs (marked by 1), hairpin loops (marked by 2), internal loops (marked by 3, 4, 5 and 6), multiloops (marked by 7) and dangling ends (marked by 8 and 9).

The free energy change of the sequence and secondary structure in Fig. 3, under the Turner99 model, is the sum of the free energy values for all structural motifs that appear in the structure, and equals -45.5 kcal/mol.

$$\Delta G = \begin{matrix} \Delta G(\text{exterior loop}) + & \sum \Delta G(\text{stackedpairs}) + \\ \sum \Delta G(\text{hairpinloops}) + & \sum \Delta G(\text{internalloops}) + \\ \sum \Delta G(\text{bulgeloops}) + & \sum \Delta G(\text{G(multi\_loops}) \end{matrix} \quad (4)$$

where the free energy for each of the structural motifs is a linear function of the free energy parameters for the aforementioned features. If we denote the sequence by x, the secondary structure by y, the parameters of the model by a vector Θ, and the number of times a feature i occurs in y by $c_i(x,y)$, then the energy function of the Turner99 model is linear in the parameters, as previously given in (3).

III. Genetic Algorithm

A. Introduction

The Genetic Algorithm (GA) is a soft computing technique which is used for optimization and search. It is based on natural selection and genetics principles [16]. Optimization concerns with finding the best solution among all possible solution which satisfy all the constraints. With some advantages, GA has also some drawbacks: slow rate of convergence, huge calculations for generations and populations etc. some modifications are needed to overcome these drawbacks and improve the performance of GA. Traditional optimization algorithm needs a better set of initial values for the design variables. If they found it, they will converge rapidly to generate good results. The problem is only the long trial and error process in finding initial values. To overcome these difficulties, a hybrid approach is presented using GA and traditional algorithms [17].

Let's take a step-by-step look at the basic process behind a genetic algorithm, illustrated in Fig. 4.

1. Genetic algorithms begin by initializing a population of candidate solutions. This is typically done randomly to provide an even coverage of the entire search space.
2. Next, the population is evaluated by assigning a fitness value to each individual in the population. In this stage we would often want to take note of the current fittest solution, and the average fitness of the population.
3. After evaluation, the algorithm decides whether it should terminate the search depending on the termination conditions set. Usually this will be because the algorithm has reached a fixed number of generations or an adequate solution has been found.
4. If the termination condition is not met, the population goes through a selection stage in which individuals from the population are selected based on their fitness score the higher the fitness, the better chance an individual has of being selected.
5. The next stage is to apply crossover and mutation to the selected individuals. This stage is where new individuals are created for the next generation.
6. At this point the new population goes back to the evaluation step and the process starts again. We call each cycle of this loop a generation.
7. When the termination condition is finally met, the algorithm will break out of the loop and typically return its finial search results back to the user.

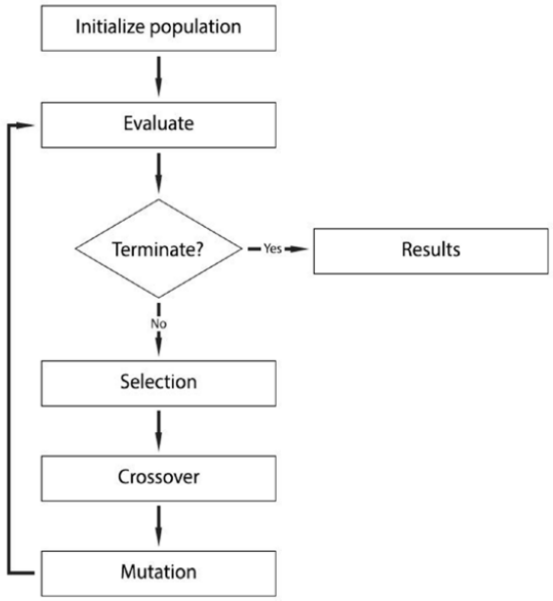

Figure 4. A general genetic algorithm process.

B. *Genetic Algorithm for RNA Secondary RNA Prediction*

- Step 1: The simplest way to apply genetic algorithms to the problem of predicting secondary structure at the initialization step is to randomly choose $y \in Y$. Then create the structure using the pair of brackets "("; ")" and ".". As follows: "(" and ")" are the links of the base canonical pairs, dots denote unmatched bases.
  Example:

$$x = CUACAAGUAUGUAG$$
$$y = (((((....))))) \quad (5)$$

- Step 2: Evaluate the y values based on the energy defined (equation (1))
- Step 3: Selection of solutions with low energy levels, while high energy solutions are eliminated, will added with childs after crossover.
- Step 4: 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 [18]. 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 [19]. However, this genetic crossover operators causes structural breakdown as defined above. For example, it may be possible to produce child puzzles with open parenthesis "(" but no closing parenthesis ")" or vice versa, or only the dot ".". This problem, authors overcome by applying fuzzy logic, will be presented in the next section.
- Step 5: Mutations, similar to crossover, also produce inappropriate structures as defined. In this paper we will not present the method of applying fuzzy logic. We will present in the next article.

IV. FUZZY LOGIC

Many, if not most books and theses on fuzzy logic start referring to the impressive success of this technique in the last decade filling whole pages with lists of successful applications. Since this is sufficiently known, it can be omitted here. I think it is more important to take a look at the challenges of the future than to exult over the success of the presence and the past. Although fuzzy methods have turned out to be very useful tools in many fields, such as control theory, expert systems, cognitive problems, signal, and image processing, robotics, to mention just a few of them, there are many questions which are still to be clarified.

In a rough outline, fuzzy systems are rule-based systems which are capable of dealing with imprecise information. Their advantage is that nearly everything inside the system can be kept interpretable for humans. Thus, prototyping can be done very easy and fast in many cases. Unfortunately, it can take a lot of time to tune all the involved parameters a problem which certainly becomes worse with increasing complexity of the system. Since the trend turns more and more to the application of fuzzy methods to very complex problems, it is necessary to deal with methods for performing this optimization (semi)automatically.

As anticipated above, we must define a mathematical framework for impreciseness which should, in some sense, be an extension of binary logic. Fuzzy logic added to the concept intermediate degrees of membership [20]. The following definition can be regarded as the basis of fuzzy logic:

Let X be an arbitrary set, the so-called universe of discourse. Then a mapping $\mu : X \to [0,1]$ is called a fuzzy subset of X. We will often abbreviate this with "fuzzy set". The set of fuzzy subsets (the fuzzy powerset) of X is denoted with $F(X) := X^{[0,1]}$. A fuzzy set μ is called normalized if and only if $\exists x \in X : \mu(x) = 1$.

The values $\mu(x)$ represent the degrees of membership to which the points x belong to the fuzzy set μ. Of course, a membership value of 0 means that an x does definitely not belong to the fuzzy set μ and a value of 1 means that x certainly belongs to μ.

Fuzzy Logic Systems Architecture in Fig. 5, it has four main parts as shown:

- **Fuzzification Module** − It transforms the system inputs, which are crisp numbers, into fuzzy sets. It splits the input signal into five steps such as

LP	x is Large Positive
MP	x is Medium Positive
S	x is Small
MN	x is Medium Negative
LN	x is Large Negative

- **Intelligence** - It stores IF-THEN rules provided by experts.
- **Rules** - It simulates the human reasoning process by making fuzzy inference on the inputs and IF-THEN rules.
- **Defuzzification Module** − It transforms the fuzzy set obtained by the inference engine into a crisp value.

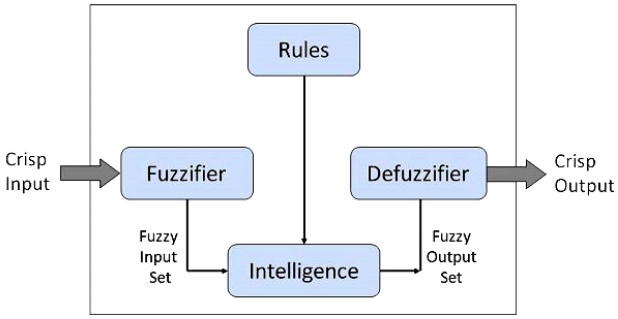

Figure 5. Block diagram of a fuzzy logic system.

V. Hybrid Genetic Algorithm with Fuzzy Logic

In this research, in this study, the authors focused application of fuzzy logic to solve the crossover operators as presented above. In genetic algorithm the crossover operator single point can lead to unexpected outcomes because of structural break, so the structural swap can be achieved by fuzzy points.

For the case of fuzzy partitions the selection of an appropriate crossing over operator is a more subtle task. Fig. 6 shows an example, where two fuzzy partitions are crossed with one point crossing over. We can summarize that the crossing over operations must be chosen depending on the given problem.

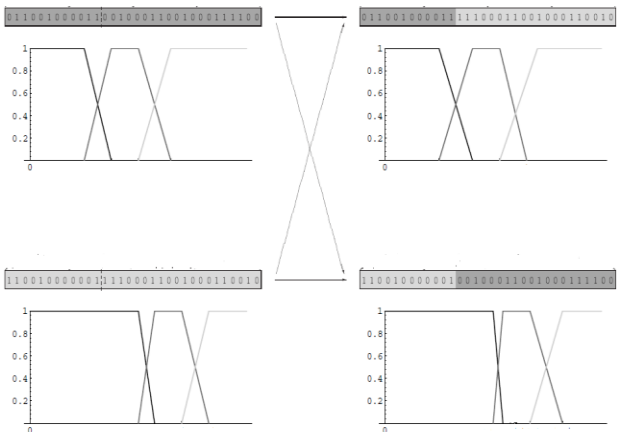

Figure 6. Crossing over one-point.

We will call it one-point crossover in the following:

Algorithm 1 One-point crossover:

```
pos  Random[1,…., n -1]
for i ← 1; pos do
    Child1[i] ← Parent2[i]
    Child2[i] ← Parent1[i]
end for
for i ←pos + 1; n do
    Child1[i] ← Parent2[i]
    Child2[i] ← Parent1[i]
end for
```

Hybrid genetic algorithms with fuzzy logic are shown in Fig. 7 as follows:

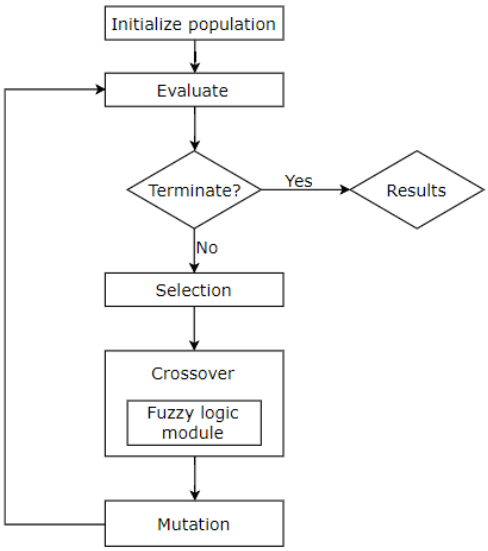

Figure 7. Hybrid genetic algorithms with fuzzy.

VI. Results and Discussion

A. The Basic Parameters of the Algorithm

```
static int AMOUNT OF INDIVIDUAL ← 100000
static List[ Individual ] population  ← new
ArrayList[ Individual ]()
static Double Pc ← 0.7
static Double Pm  ← 0.3
static float bestFreeEnergies ←  99999
static int positionOfBestFreeEnergies  ← 0
```

B. Input Sequences, Results and Comparisons

1) E.Coli 380 Bases: Sequence information: In Table I and Result and comparisons: In Table II:

TABLE I. Information of E.Coli 380 Bases Sequence

Sequence name	Information	Sequence
E.Coli 380 Bases	IDENTIFIERS: -dbEST Id:78864692 -EST name: CFSWEC34 -GenBank Acc: JZ515407 -GenBank gi: 557806030 LIBRARY:	ACUUUUAGAAGAAUU GGUUUGCUUUCAUUU UAUAUUUUUUUUGAA AGUAGACUUAUUCCU ACUUUGUUUCUUAUU UUGGGUUGGGGGUAU CAGCCGGAGCGUAUU CAGGCUGGUGUUUAU UUGUUGUUCUAUACU

	-Lib Name: LIBEST 027994 Immune responses of Coptotermes formosanus. Shiraki workers against Escherichia coli -Organism: Coptotermes Formosanus -Organ: whole body	UUGUUAGCGUCUCUU CCUUUGUUGGUUGGC AUUUUGUAUAUUUAU AGUUCAUUAGGUUCG UUAUGUUUAUUUUUG UUUCAUGGGGGUAUU GUUGGUGGUUUAUUU UAUGUUUGUAUGGUU UUGGCCUUUUUGGUU AGAAUGGCUAUAUUU UUACAACGCAGGCUU CCUACCUACUGUCCC CGGUCACGUCCCUGU UUCUGGAUCAAUGAU UUUGGAUGCAGUUUC GACCAAUCCUGUGCG CGUGU

TABLE II. ENERGY AND STRUCTURE OF E.COLI 380 BASES SEQUENCE

Genetics algorithm		Hybrid Genetics algorithm with Fuzzy logic	
Energy	Structure	Energy	Structure
-11.7	....((((((((. ....(((..........((((....... (((((((((.....)))))))))....))))............)))......))))...))))...................	-25.7	.....(((.....)))((((((((((((. ....))))))))))).... ....................(((((((...(((((((((((.....))).))))))))......(((................)))........................ ....)))))))...........

2) Bmori 498 Bases: Sequence information: In Table III and Result and comparisons: In Table IV:

TABLE III. INFORMATION OF BMORI 498 BASES SEQUENCE

Sequence name	**Information**	**Sequence**
Bmori 498 Bases	IDENTIFIERS: -dbEST Id: 45323961 -EST name: EST0545 -GenBank Acc: EL928922 -GenBank gi: 133906082 LIBRARY: -Lib Name: LIBEST 020979 ARS-CICGRU ONmgEST -Organism: Ostrinia nubilalis -Strain: bivoltine Z-pheromone strain -Sex: male and female -Organ: midgut -Develop. stage: 4th and 5th instar Larvae -Lab host: XL1 Blue -Vector: pDNRLIB	CAGAUCAUCAAGAACGACAUCGG AGUGCUGAUCACCUCCUCGCCUG UGGUGUUCACCAACCUCGUCCAA CCCAUCACUGUCUCGUAUGACUA CGCCGGUGCUGGAAUCCAGUCCA GAGCCGCUGGUUGGGGCAGAAUC AGGGCUGGCGGUCCCAUCUCCGC UCAGCUCCUCGAGUUGACCGUGA CCACCAUCUCCGGCGAUCAGUGC GUGCGUGGCGUGGCCCAGGCCUC CGUCGACUUCAACGUCGCCGCCC CACCGGUGGAACCCCACAUCGAA CUCUGCAUCAUCCACUCGCCGAA CCACGGCAUGUGUAACGGUGACU CCGGCAGCGCUCUAGUCCGCCUG GACCGCGGCACCCAGAUCGGAAU CGUGUCAUGGGGCUUCCCCUGCG CCCGCGGCGCUCCCGAUAUGUUC GUCCGAGUCAGCGCCUUCCAAGA CUGGGUCGCCCGCCACUUCGUUG CUUGAAUAAAUGACUUGAUAUGA UCGUAAAAAAAAAAA

TABLE IV. ENERGY AND STRUCTURE OF BMORI 498 BASES SEQUENCE

Genetics algorithm		Hybrid Genetics algorithm with Fuzzy logic	
Energy	Structure	Energy	Structure
–19.7	....(((((....................... .........(((((...)))))........ ..(((.............((((((.))))))......))).................... ....)))))..........	-44.2	.....................((((((((................(((.......... ....)))................ ..(((((((((((((......)))))).... (((((.(((((................... .....))))).)))))............... ....)))))))......... ....))))))))........................ ....

3) LSU 270 Bases: Sequence information: In Table V and Result and comparisons: In Table VI

TABLE V. INFORMATION OF LSU 270 BASES SEQUENCE

Sequence name	**Information**	**Sequence**
LSU 270 Bases	IDENTIFIERS: -dbEST Id: 40262660 -EST name: MVE00007287 -GenBank Acc: EC730822 -GenBank gi: 110044939 - Database: TBestDB LIBRARY: - Lib Name: LIBEST 019927 Mesostigma viride Regular library 2 - Organism: Mesostigma viride	GAAGGACGCACCGCUGGUGUACC AGUUAUCGUGCCAACGGUAAACG CUGGGUAGCCAUGUGCGGAGCGG AUAACUGCUGAAAGCAUCUAAGU AGGAAGCCCACCUCAAGAUGAGU GCUCUUUCCGUAAAAAAAACCAA AAUGGUAAAAAAAACGGUUAAGG UCACGGCAAGACGAGCCGUUUAU UAGGUGUCAAGUGGAAGUACAGC AAUGUAUGCAGCUGAGACAUCCU AACAGACCGAGGAAUUAGACCCA AGAAAGAAAUCGCUAUG

TABLE VI. ENERGY AND STRUCTURE OF LSU 270 BASES SEQUENCE

Genetics algorithm		Hybrid Genetics algorithm with Fuzzy logic	
Energy	Structure	Energy	Structure
-17.7	....((((.............((((...........(((... .........((((((((... (((..)))......))))))))............))).. ....))))............))))...................	-23.0	............((((... ((((.(((........)))))))..(((........... ..)))............... ....((((((((...(((.......))).))))))))............ ....))))...... ....

VII. Conclusion

When hybrid genetic algorithm with fuzzy logic allows apply predict RNA secondary structure for better results. Thermodynamic models in simple genetic algorithm then calculates the free energy of the structure is made better when applying DPA to model complex thermodynamics. Especially with the use of genetic algorithm with fuzzy logic hybrid is finding the optimal secondary structure is better. We will build fuzzy functions and fuzzy sets when combined with genetic algorithms to apply to the problem of predicting RNA secondary structure to achieve optimal structure.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The contributions of the authors in this paper are as follows: Binh Doan Duy, responsible for the entire content as well as the scientific content of the paper; Minh Tuan Pham, providing information on information technology and algorithms; Long Dang Duc, providing information on molecular biology; Hoan Dau Manh, provides data about the problem for the paper.

Acknowledgment

Binh Doan Duy would like to acknowledge the support of the leaders of the University of Danang - University of Science and Education, Da Nang, Vietnam, and leaders of the Faculty of Information Technology. Please acknowledge the support of lecturers from the IT Faculty, University of Science and Technology, University of Danang. The authors are grateful for the invaluable comments from the anonymous reviewers, which have strengthened this manuscript.

References

[1] J. A. Doudna and T. R. Cech, "The chemical repertoire of natural ribozymes," *Nature*, vol. 418, no. 6894, pp. 222-228, 2002.
[2] J. L. Hansen, T. M. Schmeing, P. B. Moore, and T. A. Steitz, "Structural insights into peptide bond formation," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 11670-11675, 2002.
[3] J. Bachellerie, J. Cavaill, and A. Httenhofer, "The expanding snorna world," *Biochimie*, vol. 84, no. 8, pp. 775-790(16), August 2002.
[4] M. Zuker and D. Sanko, "RNA secondary structures and their prediction," *Mathematical Bioscience*, vol. 46, pp. 591-621, 1984.
[5] H. Tsang and K. Wiese, "Sarna-predict: A study of RNA secondary structure prediction using different annealing schedules," in *Proc. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2007, pp. 239-246.
[6] M. Neethling and A. Engelbrecht, "Determining RNA secondary structure using set-based particle swarm optimization," in *Proc. IEEE Congress on Evolutionary Computation*, 2006, pp. 1670-1677.
[7] I. Tinoco and C. Bustamante, "How RNA folds," *J. Mol. Biol.*, vol. 293, no. 2, pp. 271-281, 1999.
[8] D. Mathews, J. Sabina, M. Zuker, and D. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *J. Mol. Biol.*, vol. 288, no. 5, pp. 911-940, 1999.
[9] J. Zhu and R. Wartell, "The relative stabilities of base pair stacking interactions and single mismatches in long RNA measured by temperature gradient gel electrophoresis," *Biochemistry*, vol. 36, no. 49, pp. 15326-15335, 1997.
[10] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Res.*, vol. 31, pp. 3406-3415, 2003.
[11] D. Mathews, "Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization," *RNA*, vol. 10, pp. 1178-1190, 2004.
[12] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatsh. Chem.*, vol. 125, pp. 167-188, 1994.
[13] I. Muller, *A History of Thermodynamics - The Doctrine of Energy and Entropy*, Springer, 2007.
[14] Z. J. Lu, H. D. Turner, and H. D. Mathews, "A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation," *Nucleic Acids Research*, vol. 34, no. 17, pp. 4912-4924, 2006.
[15] D. Mathews, J. Sabina, M. Zuker, and D. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *J. Mol. Biol.*, vol. 288, no. 5, pp. 911-940, 1999.
[16] R. L. Haupt and S. E. Haupt, *Practical Genetic Algorithms*, second ed., Willy Interscience, a John Willy and sons, INC., 2004, pp. 18-24.
[17] P. Guo, X. Wang, and Y. Han, "The enhanced genetic algorithms for optimisation design," in *Proc. 3rd International Conference on Biomedical Engineering and Informatics*, 2010.
[18] B. A. Shapiro and J. Navetta, "A massively parallel genetic algorithm for RNA secondary structure prediction," *J. Supercomput*, vol. 8, pp. 195-207, 1994.
[19] M. Zukar, "Mfold web server for nucleic acid folding and hybridization prediction," *Nuc. Acid. Res.*, vol. 31, no. 13, pp. 3406-3415, 2003.
[20] L. A. Zadeh, "Fuzzy sets," *Information Control*, vol. 8, pp. 338-353, 1965.

Binh Doan Duy is with Faculty of Information Technology, Danang University of Science and Education, the University of Danang, Vietnam. He received his B.A. from Faculty of Informatics, Hue University of Education, the University of Hue, Vietnam in 1998; M.A. from IT Faculty, Danang University of Science and Technology, the University of Danang, Vietnam in 2009. Now he is PhD Candidate at IT Faculty, Danang University of Science and Technology, the University of Danang, Vietnam. Since 1998, he is Lecturer at Faculty of Information Technology, Danang University of Science and Education, the University of Danang, Vietnam.
His main works include:
D. D. Binh, P. M. Tuan., D. D. Long, and N. H. Danh, "RNA secondary structure prediction by a combination of genetic algorithm with fuzzy logic," in *Proc. National Conference on Fundamental and Applied IT Research*, 2018, vol. 11, pp. 110-119.
D. D. Binh, P. M. Tuan, and D. D. Long, "Improved genetic algorithms and the application of predictive RNA secondary structures," in *Proc. National Conference on Fundamental and Applied IT Research*, 2017, vol. 10, pp. 54-67.
D. D. Binh, P. M. Tuan, and D. D. Long, "Evaluation and comparison of performing algorithms of RNA secondary structure," *Hue University Journal of Science*, vol. 121, pp. 5-19, 2016.

Minh Tuan Pham is with IT Faculty, Danang University of Science and Technology, the University of Danang, Vietnam. He received his Ph.D. from Department of Computational Science and Engineering, Graduate School of Engineering, Nagoya University in 2012. During 2012-2014, he was Lecturer at IT Faculty, Danang University of Science and Technology, the University of Danang, Vietnam. Since 2015, he is Head of Department of Computer Networks at IT Faculty, Danang University of Science and Technology, the University of Danang, Vietnam.

His main works include:

M. T. Pham and T. B. Nguyen, "The DOMJudge based online judge system with plagiarism detection," in *Proc. IEEE-RIVF International Conference on Computing and Communication Technologies*, 2019.

N. H. V. Nguyen, M. T. Pham, N. D. Ung, and K. Tachibana, "Human activity recognition based on weighted sum method and combination of feature extraction methods," *International Journal of Intelligent Information Systems*, vol. 7, no. 1, p. 9, 2018.

M. T. Pham and K. Tachibana, "A conformal geometric algebra based clustering method and its applications," *Advances in Applied Clifford Algebras*, vol. 26, no. 3, pp. 1013-1032, 2016.

Long Dang Duc is with VN-UK Institute for Research & Executive Education, the University of Danang, Vietnam. He received his Ph.D. from Department of Chemistry, University of Massachusetts- Lowell in 2003. During 2004-2009, he was Researcher at Wadsworth Medical Research Center, New York, USA. In 2010-2015, he was Head of Department of Biotechnology at Faculty of Chemistry, Danang University of Science and Technology, the University of Danang. Since 2015, he is Head of Research and International Office, VN-UK Institute for Research & Executive Education, the University of Danang.

His main works include:

W. Rennie, S. Kanoria, C. Liu, B. Mallick, D. Long, A. Wolenc, C, S. Carmack, J. Lu, and Y. Ding, "STarMirDB: A database of microRNA binding sites," *RNA Biol.*, vol. 13, no. 6, pp. 554-560, 2016.

W. Rennie, C. Liu, C. S. Carmack, A. Wolenc, S. Kanoria, J. Lu, D. Long, and Y. Ding, "STarMir: A web server for prediction of microRNA binding sites," *Nucleic Acids Res.*, vol. 42, pp. 114-118, 2014.

C. Liu, B. Mallick, D. Long, W. A. Rennie, A. Wolenc, C. S. Carmack, and Y. Ding, "CLIP-based prediction of mammalian microRNA binding sites," *Nucleic Acids Res.*, vol. 41, no. 14, p. e138, 2013.

D. Long, R. Lee, P. Williams, C. Y. Chan, V. Ambros, and Y. Ding, "Potent effect of target structure on microRNA function," *Nat. Struct. Mol. Biol.*, vol. 14, no. 4, pp. 287-294, 2007.

Hoan Dau Manh is with Learning and Research Center, Quang Binh University, Vietnam. He received his B.A. in Informatics in Hue pedagogical University in Vietnam in 1998; M.A. in Informatics in Hue University of Sciences in Vietnam in 2004; Ph.D. in Computer Science and Technology in Wuhan University of Technology in 2015.

His main works include:

M. D. Hoan and X. Ning, "The effectiveness of using methods two-stage for cross-domain sentiment classification," *Journal Computer Modelling And New Technologies*, 2014

M. D. Hoan, X. Ning, and K. T. Tung, "A survey of using weakly supervised and semi-supervised for cross-domain sentiment classification," *Advanced Materials Research*, vol. 905, pp. 637-641, 2014.

M. D. Hoan, "Feature selection using singular value decomposition and orthogonal centroid feature selection for text classification," *International Journal of Research in Engineering and Technology*, 2016.