DOI: 10.15625/vap.2020.00143

A NEW METHOD OF RNA SECONDARY STRUCTURE PREDICTION BASED ON GENETICS ALGORITHMS AND MACHINE LEARNING

Doan Duy Binh[1], Pham Minh Tuan[2] and Dang Duc Long[3]

[1]The University of Da Nang, University of Science and Education, 459 Ton Duc Thang street, Lien Chieu district, Da Nang city, Vietnam,
[2]The University of Da Nang, University of Science and Technology, 54 Nguyen Luong Bang street, Lien Chieu district, Da Nang city, Vietnam,
[3]The University of Da Nang, VN-UK Institute for Research & Executive Education, 158A Le Loi street, Hai Chau district, Da Nang city, Vietnam,
ddbinh@ued.udn.vn, pmtuan@dut.udn.vn, long.dang@vnuk.edu.vn

ABSTRACT: *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. Thus, it is not very practical to solve the problem of long sequences with dynamic programming. In this paper, we propose a novel RNA secondary structure prediction algorithm using a neural network model combined with genetics algorithms to improve the accuracy with large-scale RNA sequence and structure data. We analyze current experimental RNA sequences and structure data to construct a deep network model, and then we extract implicit features of an effective classification from large-scale data to predict the pairing probability of each base in an RNA sequence. For the obtained probabilities of RNA sequence base pairing, an enhanced genetic algorithm is applied to obtain the optimal RNA secondary structure. Results indicate that our proposed method is superior to the common RNA secondary structure prediction algorithms. Based on the characteristics of deep learning algorithm, it can be inferred that the method proposed in this paper has higher prediction success rate when compared with other algorithms, which will be needed as the amount of real RNA structure data increases in the future.*

Keywords: *Neural Network, Genetic Algorithm; Machine Learning; RNA Secondary Structure; Base Pairing; Minimum Free Energy; Long Short-Term Memory.*

I. INTRODUCTION

RNA molecules are integral components of the cellular machinery for protein synthesis and transport, transcriptional regulation, chromosome replication, RNA processing and modification, and other fundamental biological functions [1], [2]. RNA secondary structure is represented by a list of the nucleotide bases paired by hydrogen bonding within its nucleotide sequence. Studying the relationship between RNA function and structure and determining the form and frequency of RNA folding are important to reveal the role of RNA molecules in the life process [3], [4].

Secondary structure can be determined directly by x-ray diffraction, but this is difficult, slow, and expensive. Moreover, it is currently impossible to crystallize most RNAs. Mathematical models for prediction have therefore been developed and these have led to serial (and some parallel) computer algorithms, but these too are expensive in terms of computation time.

This macromolecule is basically composed of four fundamental molecules i.e., *Adenine (A), Cytosine (C), Guanine (G) and Uracil (U).* The molecules are same as that of DNA accept *Uracil*. DNA has *Thymine (T)* instead of *Uracil (U).* Another structural difference is that DNA is double stranded, however in most cases, RNA is single stranded. In the presence of salty water, RNA forms intra strand base-pairs, which result in the formation of secondary structure. Under appropriate conditions, the secondary structure folds back around itself to form tertiary structure of RNA. This folding process usually depends on the presence of divalent ions like magnesium ions and on the temperature.

Until now, much progress has been made in the computational simulation of RNA secondary structure prediction. Dynamic programming is one of the old and widely accepted techniques. This method of secondary structure prediction was first proposed by Waterman [5], Waterman and Smith [6] and Nussinov [7]. The drawback of this method of prediction is its computational time. The behavior of dynamic programming algorithms is found to be of $Ø(n^4)$, which is too slow to be effective, for bigger sequences, since the behavior is exponential. Several attempts to modify the dynamic programming algorithms have been made and considered to be successful. Another method of determining the secondary structure of RNA is the comparative method, which works simultaneously with more than one sequence in order find an identical structure. Sanko [8] extended the dynamic programming approach by folding and aligning multiple sequences to generate a phylogenetic tree for secondary structure prediction. The Zuker algorithm, implemented in the programs MFOLD [9] and ViennaRNA [10], is an efficient dynamic programming algorithm for identifying the globally minimal energy structure for a sequence, as defined by such a thermodynamic model [11], [12]. The Zuker algorithm requires $Ø(n^3)$, time and $Ø(n^2)$, space for a sequence of length N.

Unfortunately, these methods can predict only the structure of an RNA sequence with length no more than 200 in acceptable time. Corpet and Michot designed a heuristic algorithm to identify which portions of two sequences can be aligned without the structure information, and others portions are aligned by using a specialized dynamic programming algorithm [13]. This method cannot predict structures with pseudoknots. Notredame et al. used a genetic algorithm (GA)

to achieve the optimization of sequence alignment. It can solve the problem with pseudoknots and possibly predict the structure of an RNA sequence with length more than 2000 [14].

Artificial intelligence methods have been applied in many fields. At present, there have been some artificial intelligence learning algorithms such as the genetic algorithm [15], neural network algorithm [16], support vector machine algorithm, and other methods to predict the secondary structure of RNA. All achieved good results. However, all these methods are based on small samples, and the prediction accuracy is low for single-class data samples. With the development of computer technology, deep learning methods have emerged in the field of artificial intelligence, which can effectively improve the accuracy of prediction. Deep learning methods can extract effective and implicit features through deep-seated networks in large-scale data and use these features to construct effective prediction models. At present, deep learning methods have made great breakthroughs in the field of protein secondary structure prediction [17]. However, compared with secondary structure prediction of proteins, RNA secondary structure prediction is more complicated and difficult since each pair of bases on the RNA needs to correspond to another base in the chain even though each amino acid of a protein is not related to other amino acids in the chain during structure prediction.

This paper proposes a novel computational method that combines deep learning with genetic algorithm to predict RNA secondary structure prediction, which can effectively solve the problems above. Compared with the current mainstream algorithms, our method has better results.

The organization of this paper is as follows. In Section 2, we shall introduce the RNA secondary structure prediction problem is introduced. Our methods are described in Section 3. The results of the experiments described in Section 4. Conclusions about the theory and experiments are offered in Section 5.

II. SECONDARY STRUCTURE OF RNA

A. The basic concept of RNA

The Bioinformatics of *Ribonucleic Acids (RNAs)* represent a very natural realm of application, and a source of constant inspiration, for ensemble analyses. RNA constitutes a category of biomolecules, abstracted as sequence of nucleotides *Adenine (A)*, *Cytosine (C)*, *Guanine (G)* and *Uracil (U)*, initially transcribed from a DNA template and further processed before reaching their cellular environment. They can form stable complexes with proteins, but also DNA and other RNAs, allowing them to regulate genetic expression. They can also perform enzymatic functions, i.e. process other molecules (or themselves), act as biosensors by undergoing conformational changes upon binding with small metabolites, and store the entire genetic material of certain viruses (e.g. HIV, SARS, 2019 COVID).

Their dual capacity to store and process information is unmatched, leading current theories in evolution to consider RNA as the most likely candidate at the origin of life [18]. Such a versatility, illustrated in Figure 1 Figure 1, not only stems from the combinatorial nature of RNA sequences but also from its capacity to adopt one or several well-defined structures, driving the specificity of its interactions with other actors of the cellular world.

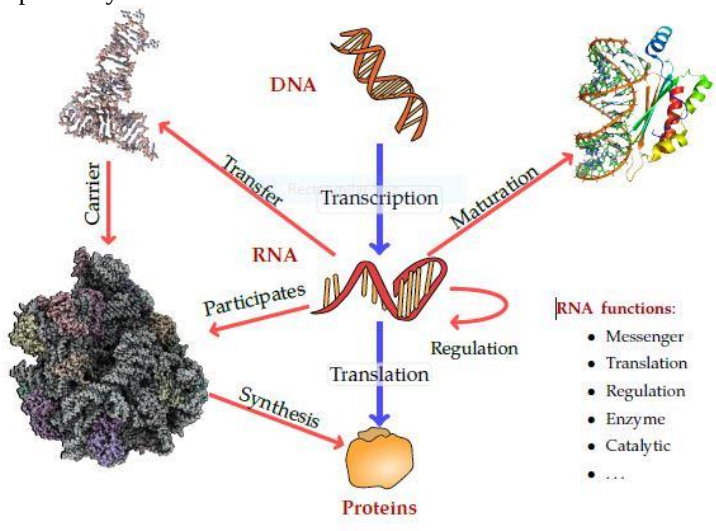

Figure 1. The basic RNA functions

B. RNA secondary structure

A RNA secondary structure of size n represents the outcome of a folding process, and focuses on a subset of base-pairs, mediated by hydrogen bonds. For essentially computational reasons [19],[20],[21] this definition forbids crossing base pairs, also called pseudoknots due to their ability to induce complex topologies [22]. Moreover, any nucleotide can only be involved in a single base pair, since additional partners would involve non-canonical edges [23]. Finally, most definitions rule out base pairs between proximate positions, due to steric effects inducing geometric constraints, leading to a minimal number Ø of unpaired positions between paired positions.

A molecule of RNA consists of a long chain of subunits, called ribonucleotides. Each ribonucleotide contains one of four possible bases: *adenine, guanine, cytosine, or uracil (abbreviated as A, G, C, U respectively).* It is this sequence of bases, known as the *primary structure* of the RNA, that distinguishes one RNA from another.

There are two basic problems encountered in the prediction approach. First is the need for accurate measures of the free energies of the various possible substructural components - of individual base pairs as well as stems, loops, and bulges. Second, the space of possible secondary structures for a given sequence is extremely large; a systematic search through all possible configurations for a minimum-energy structure can be prohibitively slow even on fast computers.

The secondary structure can be drawn in a variety of ways, as illustrated by Figure 2, many of which being supported by our popular software VARNA, developed in collaboration with Kevin Darty and Alain Denise [24].

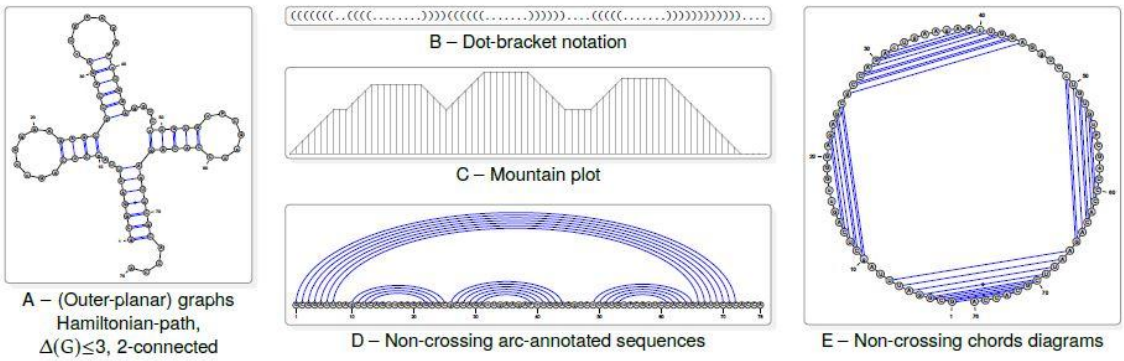

Figure 2. Various representations for RNA secondary structures

The stability of a secondary structure is assessed using a free-energy model. The most popular such model is the Turner nearest-neighbor free-energy model [25], which associates experimentally-determined free-energy contributions to structural motifs, called loops (see). The energy of any given secondary structure is then additively defined, i.e. obtained by summing the contributions of the various loops appearing in the structure.

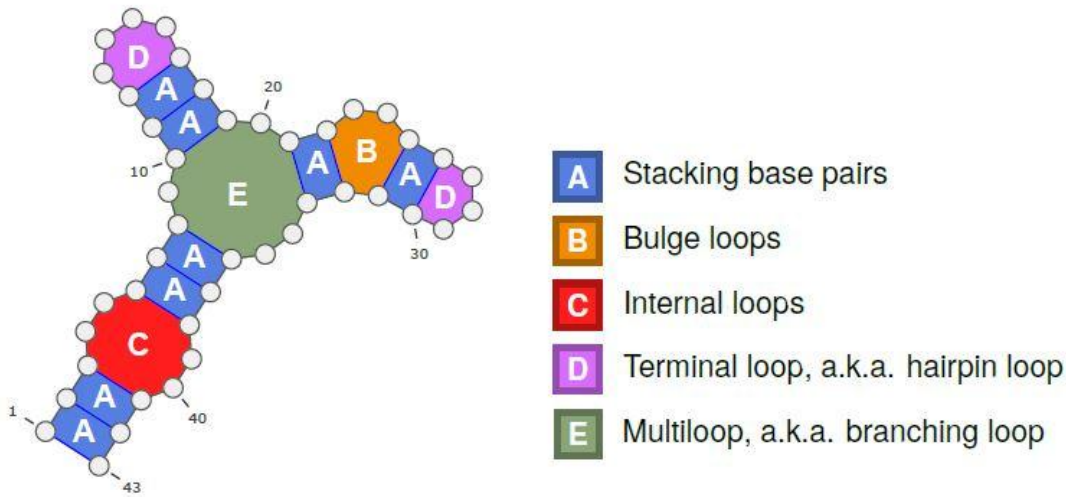

Figure 3. Loop decomposition supporting the Turner nearest neighbor model [25]. Free energies are associated to each loop type, precise topology and content in nucleotides, determined by, or extrapolations from, direct experimental measurements

We represent an RNA molecule as a sequence S of symbols: $s_1 s_2 \ldots s_n$, where s_i is one of *G*, *C*, *A*, or *U*. A subsequence of *S* may be called a "sequence" where no confusion will occur. A sequence or subsequence may also be called a "string".

Given a sequence S, we represent the secondary structure of S by the upper right triangular submatrix of an $n - by - n$ matrix A. A_{ij} is 1 i $paired(i,j)$f, i.e., (for $i < j$), if the bases at positions i and j in the sequence are paired, and is 0 otherwise. The secondary structure may then also be represented by a P of pairs, where (i,j) is in P if and only if $paired(i,j)$. A pairing itself will sometimes be referred to as $i \circ j$.

The subsequence from $i\ to\ j$ is written $[i,j]$. A subsequence is proper with respect to a secondary structure P if, for every paired element in the subsequence, its partner is also in the subsequence. If $i \circ j$ is a pair and $i < r < j$ then we sa $i \circ j$ y surrounds r. Likewise $i \circ j$ surrounds $r \circ s$ if it surrounds both r and s. (The rule against knots dictates that given $r \circ s$ if $i \circ j$ surrounds either r or s, then it surrounds both). Subsequence $[i,j]$ is closed with respect to a

structure P if $(i,j) \in P$. A pair $p \circ q$ or an element r in proper string $[i,j]$ is accessible in $[i,j]$ if it is not surrounded by any pair in $[i,j]$ except possibly $i \circ j$. It is accessible from $i \circ j$ if i and j are paired. A cycle c is a set consisting of a closing pair $i \circ j$ and all pairs $p \circ q$ and unpaired elements r accessible to it.
We can distinguish two kinds of constraints on the forming of an RNA secondary structure: *hard constraints*, *soft constraints* and *costs* re the terms often used in optimization work. Hard constraints dictate that certain kinds of pairings cannot occur at all; soft constraints are those imposed by thermodynamics upon the classes of possible structures. Hard constraints determine which structures are "legal"; soft constraints determine which structures are *optimal.*

The hard constraints are:

1. (Watson-Crick pairing): If Pcontains (i,j) then s_i *and* s_j are either *G and C*, or *C and G*, or *A and U*, or *U and A*. (This may be easily extended to include the relatively rare GU pairings.)
2. There is no overlap of pairs. If Pcontains (i,j), then it cannot contain (i,k) if $k \neq j$ or (k,j) if $k \neq i$.
3. For all i, (i,i) cannot be in P.
4. Knots are not allowed: If $h < i < j < k$, then P cannot contain both (h,j) and (i,k).
5. No sharp loops are allowed: If Pcontains (i,j), then i and j are at least 4 bases apart.

The soft constraint on possible secondary structures $P$for S is simple: S will assume the secondary structure P that has *minimum free energy*.

A secondary structure P for S can be described in a natural and unique way as composed of substructures of four kinds: loops, bulges, stacked pairs (a stack of pairs is called a *stem*), and external single-stranded regions Figure 4. The *cycles* of P are its loops, bulges, and stacked pairs. It is useful here to provide some definitions of cycles.

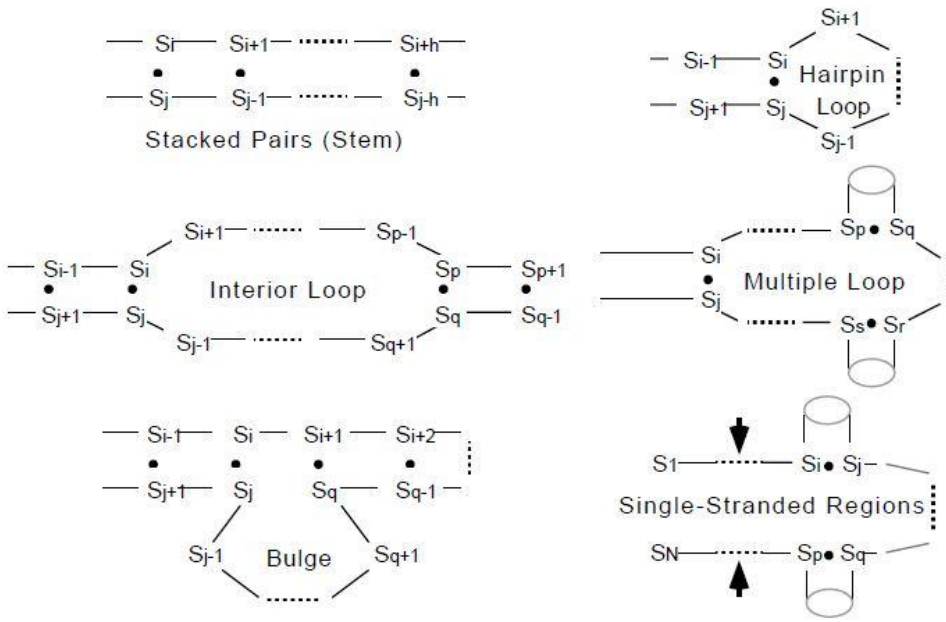

Figure 4. This figure illustrates the six basic kinds of RNA substructure. The indices S_i etc., represent base numbering, and the dots represent base pairing

1. If P contains $i \circ j$, $(i+1) \circ (j-1)$, … $(i+h) \circ (j-h)$, each of these pairs (except the last) is said to *stack* on the following pair. Two or more such consecutive pairs are called a *stacked pairs cycle*.
2. If Pcontains $i \circ j$ but none of the surrounded elements $i+1 \ldots j-1$ are paired, then the cycle is a hairpin loop. (Many molecular biologists use "hairpin" to refer to a stem with a loop of size 0 or 1 at the end, i.e., a stem with virtually no loop.)
3. If $i+1 < p < q < j-1$ and P contains $i \circ j$ and $p \circ q$, but the elements between i and p are unpaired and the elements between q and j are unpaired, then the two unpaired regions constitute an *interior loop*.
4. If P contains $i \circ j$ and $i \circ j$ surrounds two or more pairs $p \circ q, r \circ s$,… which do not surround each other, then a *multiple loop* is formed.
5. If P contains $i \circ j$ and $(i+1) \circ q$, and there are some unpaired elements between q and j, (or, symmetrically, if Pcontains $i \circ j$ and $p \circ (j-1)$ and there are unpaired elements between i and p, then these unpaired elements form a *bulge*.
6. Let r be a sequence of elements in the sequence. If r is unpaired and there is no pair in P surrounding r, then we say r is in a *single-stranded region*.

The classical (Tinoco-Uhlenbeck) approach to specifying the free energy $E(P)$of a secondary structure rests on the hypothesis that the free energy is a sum of the free energy values of Ps cycles.

$$E(P) = \sum_i E(c_i) \quad (1)$$

Even if we accept as a working assumption the equation given above, we are left with the task of specifying free energy values for the primitive substructures. For this we must turn to empirical biochemistry

III. METHODS

As a research hotspot in the field of machine learning, deep learning can mine deeper hidden features from data [26], [27]. A recurrent neural network is a sequence oriented neural-network model for deep learning that displays excellent performance in natural-language processing, image recognition, and speech recognition [28].

However, common deep neural-network models are restricted to features with a fixed shape and, therefore, cannot model RNA primary structures with variable sequence lengths. Here, we applied a long short-term memory (LSTM) network to establish a secondary structure-prediction method that is adaptable to RNA sequences of variable length. A previous study by Mathews [29] showed that a higher base-pairing probability calculated by the partition function resulted in a greater the probability of its appearance in the real structure. Therefore, the type of base and the output of its partition function was selected as the feature of the base. Additionally, we introduced a mask vector to eliminate the effect of the extended sequence on the model, which allowed the model to process variable length RNA primary sequences.

A. Long Short-Term Memory (LSTM) Network

LSTM network is a type of deep RNN model composed of LSTM units. As discussed earlier, recurrent neural network (RNN) is a deep learning network with internal feedback between neurons. These internal feedbacks enable the memorization of significant past events and incorporate past experienced. Unlike a traditional fully connected feedforward network, RNN shares parameters across all the parts of a model, so it can be generalized to sequence lengths that have not been seen during training. Figure 5 presents an example of RNN architecture that produces an output at every time step, and has recurrent connections among hidden neurons [30].

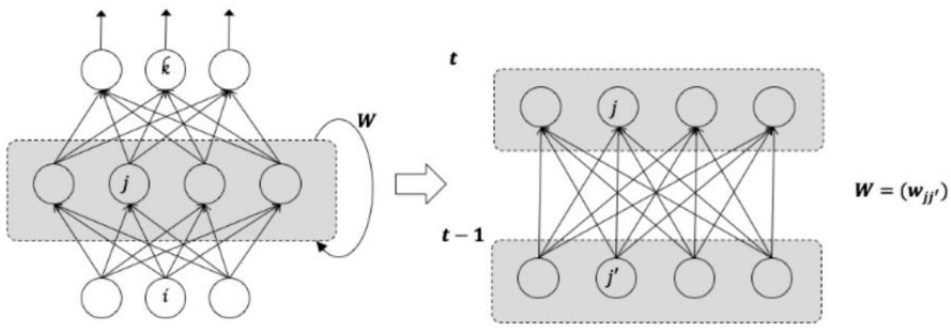

Figure 5. Basic structure of a simple recurrent neural network (RNN)

The LSTM block contains memory cell in Figure 6 and three multiplicative gating units; an input, an output, and a forget gate. There are recurrent connections between the cells, and each gate provides continuous operations for the cells. The cell is responsible for conveying "state" values over arbitrary time intervals, and each gate conducts write, read, and reset operations for the cells [31], [32].

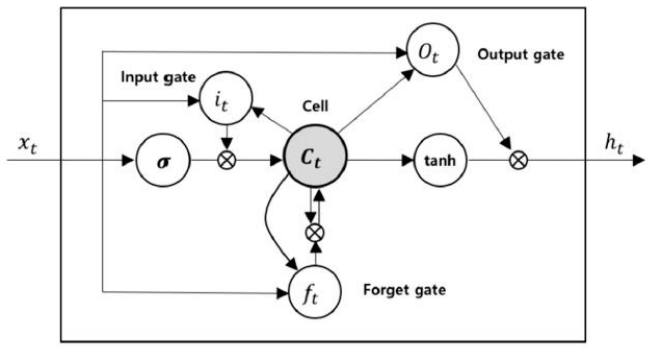

Figure 6. Long short-term memory (LSTM) cell with gating units

B. Genetic Algorithm (GA)

GA is metaheuristic and stochastic optimization algorithm inspired by the process of natural evolution [33]. They are widely used to find near-optimal solutions to optimization problems with large search spaces. The process of GA includes operators that imitate natural genetic and evolutionary principles, such as crossover and mutation. The major

feature of GA is the population of "chromosomes". Each chromosome acts as a potential solution to a target problem, and usually expressed in the form of binary strings. These chromosomes are generated randomly, and the one that provides the better solution gets more chance to reproduce.

Processing the GA can be divided into six stages: initialization, fitness calculation, termination condition check, selection, crossover, and mutation, as shown in Figure 7 [33]. In the initialization stage, a chromosome in the search space is arbitrarily selected, and then the fitness of each selected chromosome is calculated in accordance with the predefined fitness function. The fitness function is a concept used to numerically encode a chromosome's performance. In optimization algorithms, such as GA, the definition of a fitness function is a crucial factor that affects the performance. Through the process of calculating the fitness for the fitness function, only solutions with excellent performance are preserved for further reproduction processes.

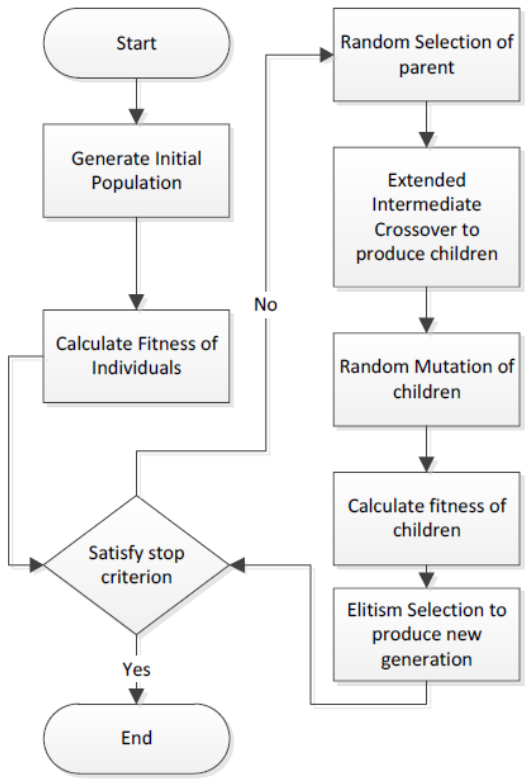

Figure 7. Basic process of a genetic algorithm (GA)

C. A Hybrid Approach to Optimization in LSTM Network with GA

Evolutionary algorithms, mostly GA, have been widely applied to neural network models, such as multi-layer perceptron (MLP) and RNN, and used in various hybrid approaches for RNA secondary structure prediction.

In this study, we propose a hybrid approach of LSTM network integrating GA to find optimal RNA secondary structure. Figure 8 depicts the flowchart of the model proposed in our work.

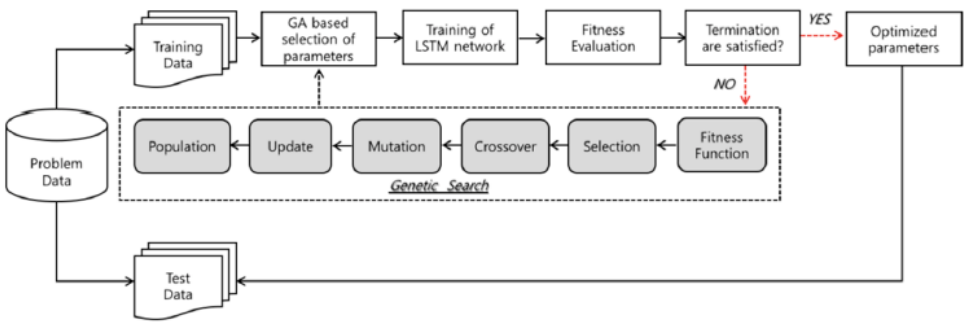

Figure 8. Flowchart of the GA–LSTM model

D. RNA secondary structure prediction

A, G, C, U are four different bases in RNA molecules, several bases are arranged order to form the primary structure of RNA [34]. The primary structure of an RNA sequence S consisting of n bases can be expressed as $S =$

$s_1 s_2 \ldots s_n$, where s_1 is the base near the $5'$ side, s_n is the base near the $3'$ side, s_i is the $i-th$ base in sequence S and $s_i \in \{A, G, C, U\}$ RNA secondary structure prediction problem, with the purpose of calculating the pairing results y_i of each base s_i in sequence S when the primary structure of S is known, is a classification problem. According to different categories of classification, RNA secondary structure prediction problems can be divided into the following categories:

(1) Two categories classification: pairing results consist of the category of paired base ($y_i = 1$) and the category of not paired base ($y_i = 0$).

(2) Three categories classification: pairing results include the category of paired base near $5'$side ($y_i = 1$), the category of paired base near $3'$side ($y_i = 2$) and the category of not paired base ($y_i = 0$).

(3) Multi-category classification: for a sequence with pseudoknots, pairing results $y_i = j$ ($j = 0,1,\ldots,n$) means the $i-th$ base is paired with the $j-th$ base if $j > 0$, otherwise the $i-th$ base is not paired.

In this article we focus on categories 1 and 3.

E. Adaptive LSTM with energy-based model

The scheme of a RNA secondary structure-prediction model based on Adaptive LSTM and energy-based filter.

- Adaptive LSTM: A mask vector was introduced to enable the model to effectively process sequences of different length and to ensure that the extended base will not affect the normal training of the model. The mask vector is used to distinguish between the original and extended parts of a sequence, and the new sequence represents the input to the LSTM network.
- Energy-based filter: As the result of translating the RNA secondary structure prediction problem into a classification problem of base pairings, there exist some conflicting pairing result in the output of LSTM. The energy-based filter is used to deal with this problem. In laws of thermodynamics, RNA structures with a lower free energy are more stable [35], so the energy based is used to randomly change the label of conflicting base pairings according to the free energy of the structure to make the structure more likely to its real structure.

  According to the Watson-Crick base complementary pairing principle [36], each base $s(i)$, can only interact with at most one other base, $s(j)$, to form one base pair $\big(s(i), s(j)\big)$ and $\{s(i), s(j)\} \in \big\{\{A, U\}, \{C, G\}, \{G, U\}\big\}$. As a result, the predicted result of $i-th$ base $y(i)$ can be reserved if the two following conditions are met:

  1 $y\big(y(i)\big) = i$

  2 $\Big(s(i), s\big(y(i)\big)\Big)$ is in $\{(A, U), (U, A), (C, G), (G, C), (G, U), (U, G)\}$

  If these conditions are not met, $y(i)$ should be set as 0 to classify the $i-th$ base as unpaired base.
  By setting all the conflicting bases to unpaired bases, it may incorrectly turn false positive samples into false negative samples, energy-based filter is improved from reducing unmatched pairs to filter pairing results by the free energy of secondary structure.

IV. RESULTS AND DISCUSSION

A. The basic parameters of the algorithm

The parameters for the genetic algorithm presented in this paper are as follows:
- Static int AMOUNT OF INDIVIDUAL 100000
- Static List [Individual] population
- Array List [Individual] ()
- Static Double Pc 0.7
- Static Double Pm 0.3
- Static float best Free Energies 99999
- Static int position of Best Free Energies 0

The dataset of this paper comes from authoritative dataset RNA STRAND. The RNA STRAND database can be used to evaluate the prediction accuracy of energy-based RNA secondary structure prediction software. RNA STRAND is publicly available at http://www.rnasoft.ca/strand.

B. Input sequences, the results and comparisons

The results in this paper are compared with the results of a study by the same group of authors shown in [37] [38].

1. E. Coli 221 Bases

Sequence information: In 0, Result and comparisons: In Table 2 and Model RNA secondary structure for final results Figure 9:

Table 1. Information of E. Coli 221 Bases sequence

Sequence name	Information	Sequence
E. Coli 221 Bases	IDENTIFIERS: - dbEST Id: 78864696 - EST name: CFSWEC38 - GenBank Acc: JZ515411 - GenBank gi: 557806034 LIBRARY: - Lib Name: LIBEST_027994 Immune responses of Coptotermes formosanus.Shiraki workers against *Escherichia coli* - Organism: *Coptotermes formosanus* - Organ: whole body	ACAUGGGGAUAAGGGCAGGCGGUGA UGCCUUGGCUCUCGGAGGCGAAGAAG GACGUGAUAAGCUGCGAUAAGCCCG CGUAGGCGCAAAUAGCCGUUAAUACC GGGGUUUCCGAAUGGGGCAACCCGCC GGGAGUAAUUCCGGCAUCUCUUGAA GAGGGAGGCGAACGUGGGGAACUGAA ACAUCUCAGUACCUGCAGGAAAAAAA

Table 2. Energy and Structure of E. Coli 221 Bases sequence

Genetics algorithm		Hybrid Genetics algorithm with Fuzzy logic		Hybrid Genetics algorithm with Machine learning	
Energy	Structure	Energy	Structure	Energy	Structure
-23.7	.((((..... .....(((((..... .....(((((((.... .....(((..)))... .....))))))))..... .....)))))).....)))).....	**-31.9**	....(((((((((.(((..... .....)))..... .....(((((((.....))))))..))).))))))).... ..... .((((((..((((((.....)))))).)))))....	**-66.53**	.....(((((.....(((((.(((((..((((.(((..(((.....))).))). .)))).)))))).(((.....))).....))))) .(((((((....(((....)))(((((((....)))))))).((((((....)))))).((....)).)))))))).....)))))....((....)).....

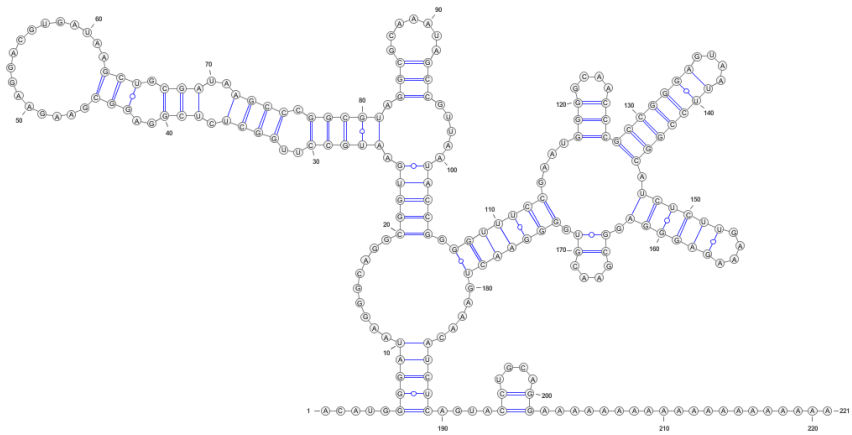

Figure 9. Secondary structure of E. Coli 221 bases

2. B. Mori 219 Bases

Sequence information: In Table 3, Result and comparisons: In Table 4 and Model RNA secondary structure for final results Figure 10:

Table 3. Information of B. Mori 219 Bases sequence

Sequence name	Information	Sequence
B. Mori 219 Bases	IDENTIFIERS: - dbEST Id: 45323636 - EST name: EST0213 - GenBank Acc: EL928597 - GenBank gi: 133905746 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	ACCACGGCAUGUGCAACGGUGA CUCCGGCAGCGCUCUAGUCCGC CUGGACCGCGGCACCCAGAUCG GAAUCGUGUCGUGGGGCUUCCC CUGCGCCCGCGGCGCUCCCGAU AUGUUCGUCCGAGUCAGCGCCU UCCAAGACUGGGUCGCCCGCCA CUUCGUUGCUUGAAUAAAUGAC UUGAUAUGAUCGUCAAAAAAAA AAAAAAAAAAAAAAAAAAAAA

Table 4. Energy and Structure of B. Mori 219 Bases sequence

Genetics algorithm		Hybrid Genetics algorithm with Fuzzy logic		Hybrid Genetics algorithm with Machine learning	
Energy	Structure	Energy	Structure	Energy	Structure
-25.8	.....(((((((..... ..((((((...(((....((((((.((((.....)))))))))).))).. ....))))))...........))))))).....	**-34.6**	.....((((.....(((.....)))...)))).(((.....((((((((((.....))))))))))..... .....((((((... .....))).)))..))).....	**-60.70**	.....((((..(((.....))) ...)))).....((((....)))).((((((((((.(((((..(((((((.((((...)))).((((((...))))))...)))))))....)))))).....((... .)).)))))).)))).))....(((((. .....)))))..(((((.....)))))......

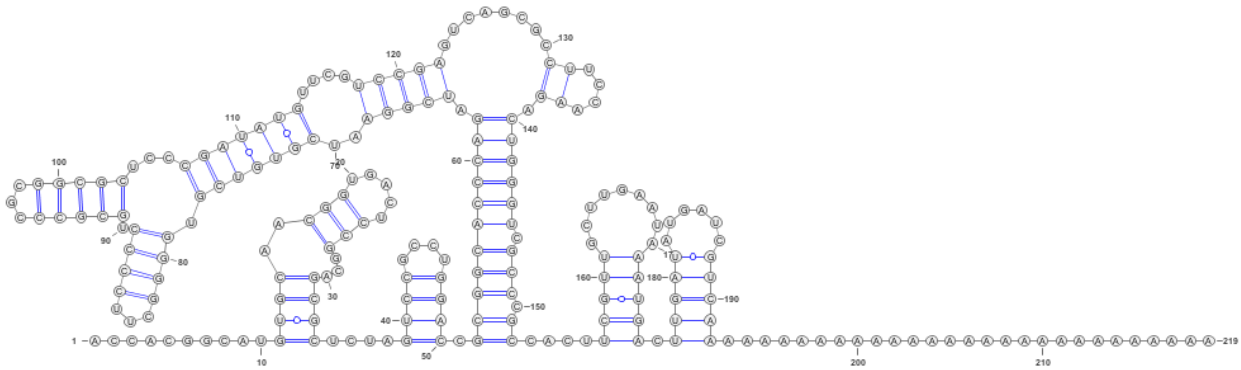

Figure 10. Secondary structure of B. Mori 219 Bases

V. CONCLUSION

It is evident that working on base-pairs reduces much of the time and complexity in predicting the secondary structure of RNA. Time complexity may be reduced further by using parallel processing, like we present by using a very simple customized architecture of Neural Network. We hybrid Genetics algorithm with machine learning allows apply predict RNA secondary structure for optimal secondary structure from the bulk of structures. This method reduces the time complexity by working on suboptimal structures rather than to let the system converge toward an optimal secondary structure, which may take a long time specifically with bigger sequences. 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 machine learning hybrid is finding the optimal secondary structure is better. This paper proposed the RNA secondary structure predicting method based on Adaptive LSTM and energy-based filter combined with genetic algorithms. This method to apply to the problem of predicting RNA secondary structure to achieve optimal structure.

REFERENCES

[1] G. Storz, “An expanding universe of noncoding rnas”, Vol. 296, pp. 1260-1263, 2002.

[2] S. Eddy, “Noncoding rna genes and the modern rna world”, Nature Reviews Genetics, Vol. 2, pp. 919-929, 2001.

[3] J.-S. Wu and Z.-H. Zhou, “Sequence-based prediction of microrna-binding residues in proteins using cost-sensitive laplacian support vector machines”, Vol. 10, pp. 752-759, 2013.

[4] W.-L. Guo and D.-S. Huang, “An efficient method to transcription factor binding sites imputation via simultaneous completion of multiple matrices with positional consistency”, Vol. 13, pp. 1827-1837, 2017.

[5] M. Waterman, “Secondary structure of single-stranded nucleic acids”, in studies on foundations and combinatorics, advances in mathematics supplementary studies, Vol. 1. ACADEMIC PRESS N.Y., pp. 167-212, 1978.

[6] M. S.Waterman, “Efficient sequence alignment algorithms”, Vol. 108, pp. 333-337, 1984.

[7] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman, “Efficient sequence alignment algorithms”, Vol. 35, pp. 68-82, 1978.

[8] D. Sanko, “Simultaneous solution of the rna folding alignment and protosequence problems”, Vol. 45, p. 810825, 1985.

[9] M. Zuker, “On finding all suboptimal foldings of an rna molecule," Vol. 244, pp. 48-52, 1989.

[10] Hofacker, W. Fontana, P. Stadler, S. Bonhoeffer, and M. Tacker, “Fast folding and comparison of rna secondary structures”, Monatshefte fuer Chemie/Chemical Monthly, Vol. 125, pp. 167-188, 1994.

[11] M. Zuker and P. Stiegler, “Optimal computer folding of large rna sequences using thermodynamics and auxiliary information”, Nucleic acids research, Vol. 9, pp. 133-48, 1981.

[12] M. Zuker and D. Sankoff, “Rna secondary structures and their prediction”, Vol. 46, pp. 591-621, 1984.

[13] F. Corpet and B. Michot, “Rnalign program: alignment of rna sequences using both primary and secondary structures”, Computer applications in the biosciences CABIOS, Vol. 10, pp. 389-399, 1994.

[14] C. Notredame, E. O'Brien, and D. Higgins, "Raga: Rna sequence alignment by genetic algorithm", Nucleic acids research, Vol. 25, No. 22, p. 4570-4580, 1997.
[15] Y.-J. Hu, "Gprm: A genetic programming approach to finding common rna secondary structure elements", Vol. 31, No. 13, pp. 3446-3469, 08 2003.
[16] X. Zhang, Z. Deng, and D. Song, "Neural network approach to predict rna secondary structures", Vol. 46, pp. 1793-1796, 10 2006.
[17] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein secondary structure prediction using deep convolutional neural_elds", Scientific reports, Vol. 6, 01 2016.
[18] P. G. Higgs and N. Lehman, "The RNA world: molecular cooperation at the origins of life", Nature Reviews Genetics, Vol. 16, pp. 7-17, 2015.
[19] T. Akutsu, "Dynamic programming algorithms for rna secondary structure prediction with pseudoknots * 1,* 2", Discrete Applied Mathematics, Vol. 104, pp. 45-62, 08 2000.
[20] R. Lyngs and C. Pedersen, "Rna pseudoknot prediction in energy-based models," Journal of computational biology : a journal of computational molecular cell biology, Vol. 7, no. 3-4, p.409427, 2000.
[21] S. Hammer, Y. Ponty, W. Wang, and S. Will, "Fixed-parameter tractable sampling for RNA design with multiple target structures", BMC Bioinformatics, Vol. 20, p. 209, 04/2019.
[22] Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert, "Prediction and statistics of pseudoknots in rna structures using exactly clustered stochastic simulations", Proceedings of the National Academy of Sciences of the United States of America, Vol. 100, No. 26, p. 15310-15315, December 2003.
[23] N. Leontis and E. Westhof, "Geometric nomenclature and classification of rna base pairs", RNA (New York, N.Y.), Vol. 7, No. 4, p. 499512, April 2001.
[24] K. Darty, A. Denise, and Y. Ponty, "VARNA: Interactive drawing and editing of the RNA secondary structure", Bioinformatics, Vol. 25, No. 15, pp. 1974-1975, 04 2009.
[25] D. Turner and D. Mathews, "Nndb: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure", Nucleic acids research, Vol. 38, pp. D280-2, 10/2009.
[26] H.Wu, K.Wang, L. Lu, Y. Xue, Q. Lyu, and M. Jiang, "Deep conditional random eld approach to transmembrane topology prediction and application to gpcr three-dimensional structure modeling", Vol. 14, pp. 1106-1114, 2017.
[27] L. Hongshun, Y. Hua, and G. Xiujun, "A deep learning model for predicting rna-binding proteins only from primary sequences", Vol. 55, pp. 93-101, 2018.
[28] Z. Shen, W. Bao, and D.-S. Huang, "Recurrent neural network for predicting transcription factor binding sites", Vol. 8, No. 15270, 2018.
[29] D. H. Mathews, "Using an rna secondary structure partition function to determine confidence in base pairs predicted by free energy minimization", Vol. 10, pp. 11781190, 2004.
[30] Goodfellow, Y. Bengio, and Y. Courville, A.and Bengio, "Deep Learning". MIT Press: Cambridge, MA, USA, 2016.
[31] F. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget continual prediction with lstm", Vol. 12, pp. 2451-2471, 1999.
[32] J. Holland, Adaptation in Natural and Artificial Systems. University of Michigan Press: Ann Arbor, MI, USA, 1975.
[33] G.Armano, M.Marchesi, and A.Murru, "Learning to forget: continual prediction with lstm", Vol. 12, pp. 2451-2471, 1999.
[34] J. Yang, R. Jang, Y. Zhang, and H.-B. Shen, "High-accuracy prediction of transmembrane inter-helix contacts and application to GPCR 3d structure modeling", Vol. 29, pp. 2579-2587, 2013.
[35] D. H. Mathews, "Using the rna structure software package to predict conserved rna structures", Vol. 46, pp. 1-22, 2014.
[36] H. Wu, C. Cao, X. Xia, and Q. Lu, "Unified deep learning architecture for modeling biology sequence", Vol. 15, pp. 1445-1452, 2018.
[37] Doan Duy Binh, Pham Minh Tuan, Dang Duc Long, "Cải tiến thuật toán di truyền và áp dụng dự đoán cấu trúc bậc hai RNA", Kỷ yếu hội nghị Quốc gia lần thứ X về nghiên cứu cơ bản và ứng dụng công nghệ thông tin (FAIR), pp. 54-67, 2017.
[38] Doan Duy Binh, Pham Minh Tuan, Dang Duc Long, and Nguyen Huu Danh, "Dự đoán cấu trúc bậc hai rna bằng sự kết hợp thuật toán di truyền và logic mờ", Kỷ yếu hội nghị Quốc gia lần thứ XI về nghiên cứu cơ bản và ứng dụng công nghệ thông tin (FAIR), pp. 110-119, 2018.

MỘT PHƯƠNG PHÁP MỚI TRONG DỰ BÁO CẤU TRÚC BẬC HAI CỦA RNA DỰA TRÊN THUẬT TOÁN DI TRUYỀN VÀ HỌC MÁY

Đoàn Duy Bình, Phạm Minh Tuấn và Đặng Đức Long

TÓM TẮT: *Nhiều phương pháp có thể được sử dụng để dự đoán cấu trúc bậc hai của phân tử RNA. Một trong những phương pháp đó là phương pháp quy hoạch động. Tuy nhiên, cách tiếp cận quy hoạch động thường mất quá nhiều thời gian. Vì vậy, nó không thực tế lắm để giải quyết bài toán với chuỗi dài với quy hoạch động. Trong bài báo này, chúng tôi đề xuất một thuật toán dự đoán cấu trúc bậc hai RNA mới bằng cách sử dụng mô hình mạng nơron kết hợp với các thuật toán di truyền để cải thiện độ chính xác với dữ liệu cấu trúc và chuỗi RNA có chiều dài lớn. Chúng tôi phân tích chuỗi RNA thực nghiệm hiện tại và dữ liệu cấu trúc để xây dựng mô hình mạng sâu, sau đó chúng tôi trích xuất các đặc điểm ngầm của phân loại hiệu quả từ dữ liệu quy mô lớn để dự đoán xác suất kết cặp của mỗi base trong chuỗi RNA. Đối với xác suất thu được của sự kết cặp base trong chuỗi RNA, một thuật toán di truyền nâng cao được áp dụng để thu được cấu trúc bậc hai RNA tối ưu. Kết quả chỉ ra rằng phương pháp đề xuất của chúng tôi vượt trội hơn so với các thuật toán dự đoán cấu trúc bậc hai RNA phổ biến. Dựa trên các đặc điểm của thuật toán học sâu, có thể suy ra rằng phương pháp được đề xuất trong bài báo này có tỷ lệ dự đoán thành công cao hơn khi so sánh với các thuật toán khác, điều này sẽ cần thiết khi lượng dữ liệu cấu trúc RNA thực tăng lên trong tương lai.*