RNA Secondary Structure Prediction Using Genetic Algorithm

Mohammad Ganjtabesh[a]*, Milad Chenaghlou[a], Abbas Nowzari-Dalini[a]

[a] Department of Computer Science, School of Mathematics, Statistics, and Computer Science, College of Science, University of Tehran, Tehran, Iran

Abstract. RNA molecules play several fundamental roles in any living cell. The function of an RNA molecule is highly related to its structure. Since the experimental determination of RNA structure is very expensive and time consuming, many efforts have been focused on computational prediction of RNA secondary structure. In this paper, a genetic algorithm given by Wiese is presented and it is improved in both speed-up and accuracy. First, a method is presented to speed-up the convergence rate of the genetic algorithm using the helices that appear with high probability in the RNA secondary structure. To improve the accuracy of the algorithm, sub-maximal helices are used instead of maximal ones. Then a hill-climbing method is used to further improve the quality of the results. Finally, the results obtained by Wiese original algorithm and those obtained by our proposed improvements are presented and compared.

Keywords: RNA secondary structure; RNA folding; Genetic algorithm.

1. Introduction

Ribonucleic Acid (RNA) is an important molecule that performs a wide range of functions in biological processes [1]. It contains the genetic information of many viruses, such as HIV, and thereby regulates their functions [2, 3]. RNA is essential for the function of a cell and it is the product of gene transcription. In a process known as translation, the RNA is involved in building the proteins. In the translation process, the ribosome uses tRNA to produce proteins where three consecutive RNA nucleotides form a codon, which encodes one of the 20 amino acids [4].

An RNA sequence is made of four different nucleotides, namely Adenine (A), Cytosine (C), Guanine (G), and Uracil (U), and it tends to fold to itself and form pairs of bases by constructing hydrogen bonds between the complementary bases. There are mainly two kinds of base pairs: 1) Watson-Crick base pairs that can be formed between bases A and U as well as between C and G, and 2) Wobble base pair that can be formed between bases G and U. It is worth noting that the other kinds of interactions between nucleotides are exist in real structure, but they are usually ignored in computational methods. The set of all conformed base pairs (between the complementary bases) is called RNA secondary structure, in which any two base pairs is either nested or disjoint. When the base pairs cross each other, the resulting structure is called RNA pseudoknotted structure. The RNA secondary structure possesses various considerable properties of RNA tertiary structure [5, 6] and the biological function of an RNA is assumed to be related to its secondary structure [7]. Any RNA secondary structure is comprised of different structural components such as stems, hairpins, bulges, internal loops, multi loops, and external loop [8].

Currently, the only accurate method for determining the RNA structure is the X-ray crystallography; however, it is not only time consuming, but also expensive [9]. Therefore, many computational methods have been proposed to predict RNA secondary structures [3, 4, 10]. Although computational methods sometimes provide an approximate RNA structure, they mostly facilitate the future studies of RNA structures.

The first attempt for predicting the RNA secondary structure was given by Nussinov, who devised a dynamic programming algorithm to maximize the number of base pairs in a given RNA sequence [11 -13]. The main drawback of this method is that maximizing the number of base pairs does not essentially produce the best structure. To overcome this, several approaches have been proposed in the literature. The first approach is based on comparative sequence analysis [14]. This approach infers base pairs by determining canonical pairs that are common among multiple homologous sequences. Comparative sequence analysis is quite robust when several homologous sequences are available. However,

* Corresponding author.
E-mail address: mgtabesh@ut.ac.ir

it requires multiple homologous sequences, it can be time consuming and it may need significant insight. The second approach is based on free energy minimization [10, 15 – 17]. The main idea behind this approach is that all molecules in the nature tend to minimize their free energy. In this particular problem, each base pair (between two complementary nucleotides) lowers the free energy depending on several characteristics of the whole structure. Algorithms in this approach try to predict the specific base pairs that minimize the free energy of the whole structure. Dynamic programming algorithms given by Zucker [17 - 19], as well as genetic algorithms given by van Batenburg [20, 21] and Wiese [16] are included in this category. The third approach is based on stochastic context free grammars [9]. In this approach, each secondary structure of an RNA sequence is considered as a derivation of a stochastic context free grammar and the predicted secondary structure is the most probable derivation of the employed grammar.

In this paper, the prediction of RNA secondary structure (without any crossing base pairs) is considered. The genetic algorithm, devised by Wiese [16], to solve this problem is presented and its weak points are improved. To do this, three extensions are proposed as follows. Creating the initial population of the genetic algorithm using another genetic algorithm greatly reduces the computational time. Using sub-maximal helices, instead of maximal ones makes the prediction more accurate. Performing a local search method over the solutions further improves the quality of the results. The rest of this paper is organized as follows. In Section 2, the genetic algorithm devised by Wiese and our proposed improvements for this algorithm are discussed. Finally in Section 3, the obtained results for Wiese original algorithm as well as our proposed improvements are presented and compared.

2. Materials and methods

2.1. Genetic algorithm of Wiese

Wiese presented a genetic algorithm (GA) to predict the secondary structure of RNA molecules [16], where the chromosomes are encoded as permutations. More specifically, the proposed algorithm predicts the specific canonical base pairs that form hydrogen bonds and build helices, also known as stems. In his algorithm, all possible helices are found in the first step. Then, an initial population of permutations over the helices is created. Each permutation has a length equal to the number of found helices. Then the selection, crossover and mutation operations are performed one after another during the main loop of this algorithm. To compute the fitness value corresponding to each chromosome, it is decoded to represent a valid structure. To this ends, the order in which helices are appeared in a chromosome (permutation) is the order in which they are picked by the decoder to be inserted into the final structure. Here, those helices that overlap with any previously selected helices are ignored. A formal description of this decoding procedure is provided in Algorithm 1. Representing the chromosomes as permutations has many distinguished benefits. More importantly, it is simple and easy to implement, as well as many crossover operations are available for this representation.

Algorithm 1: Permutation to structure

```
Data: H = (h_1, h_2, ..., h_n) is a permutation of all helices representing a
      chromosome ;
S ← ∅;
for i ⟵ 1 to n do
    if h_i does not overlap with any helices in S then
        S ← S ∪ {h_i};
```

2.2. Proposed improvements

Although Wiese algorithm was a step forward in heuristic algorithms proposed for RNA secondary structure prediction, it has several disadvantages as well. It has slow convergence rate to find a good solution due to the huge number of possible helices. On the other hand, the helices are appeared randomly in the chromosomes and no mechanism is employed to take into account that longer helices appear with more probability in RNA secondary structures. That's why this algorithm has rather slow convergence rate.

It is also observed that in some, even short, RNA sequences the algorithm does not converge to the best solution. In these cases, some base pairs do not lower the free energy of the structure. For example, some base pairs in either end of a helix make the whole structure unstable. Wiese genetic algorithm does not consider these kinds of situations.

In this paper, three extensions are introduced to overcome the above mentioned disadvantages. These extensions are provided in the following sections.

2.2.1. Generating initial population

In order to speed up the convergence rate of the algorithm, some biological facts are considered. It can be inferred from the thermodynamic laws, that longer helices are more stable than shorter ones and they have more chance to be in candidate structures. In order to give more chance to those longer helices to be appeared in candidate structures, an initial genetic algorithm is used to generate the initial population for the main algorithm, instead if generating it randomly. In this initial genetic algorithm, only the longer helices are used to increase the probability of the presence of them in the candidate structures. The population of the last iteration of this initial algorithm forms the backbone for the chromosomes in the main genetic algorithm. Once the initial population of the main genetic algorithm is obtained, some random permutations are also added to it in order to maintain the dispersion in the search space. Using this mechanism, longer helices have more chance to be appeared in the constructed RNA structures. The generated chromosomes in the last iteration of the initial genetic algorithm should be extended to appropriate chromosomes to be used as initial population in the main genetic algorithm. To this ends, Let A be a chromosome from the last population of the initial genetic algorithm. To extend A to a chromosome A0, all the remaining helices are randomly ordered and added to the end of A.

This also helps to maintain dispersion in the search space. The pseudocode for creating the initial population is given in Algorithm 2.

2.2.2. Using sub-maximal helices instead of maximal ones

In order to increase the flexibility of the algorithm to predict more accurate structures, we use sub-maximal helices instead of maximal ones. The maximal and sub-maximal helices are defined as follow:

- Maximal helix: a helix, say A, is maximal if there is no longer helix, say B, containing all base pairs of A.
- Sub-maximal helix: every set of 3 or more consecutive base pairs, which is not a maximal helix, is a sub-maximal helix.

In Wiese genetic algorithm, maximal helices are used and therefore the base pairs that make the whole structure unstable cannot be modified. Here, we used sub-maximal helices to increase the flexibility of the algorithm. Although using sub-maximal helices increases the number of possible helices (consequently the length of the permutations) and slows down the genetic algorithm, but the obtained structures in this way are more accurate.

Algorithm 2: Generate_Initial_Population (RNA sequence R, All helices H, $ratio$)

```
begin
    ℓ ← ration × |H|
    P ← Create initial population where each chromosome is of length ℓ
    while stopping criteria is not reached do
        C ← crossover(P)
        C ← mutation(C)
        P ← Select_New_Population(P ∪ C)
    foreach p_i ∈ P do
        rH ← H − p_i
        rH ← Permute(rH)
        p_i ← p_i + rH
    return (P)
```

2.2.3. Using a local search to lower the free energy in a targeted manner

When using sub-maximal helices, it is observed that a base pair in either end of a helix may produce unstable structure. Although these kinds of base pairs will be finally removed in genetic algorithm, either by crossover or mutation operations [22, 23], but it may require many iterations. Therefore, we used a local search method, i.e. hill climbing, to consider the effect of existing or absence of these base pairs in free energy of the structure. To do this, each helix in the chromosome is picked and the structure obtained by removing a single base pair from either end of the helix or possibly by adding a base pair to either end of it as well as the free energy of them are calculated. Then the best structure is selected to be in the current population. Since applying the hill climbing method over all chromosomes in the current population is very time consuming, therefore we only apply it over the small portion of the best structures in the current population.

3. Results and discussion

In order to compare the results of our proposed genetic algorithm with those obtained by Wiese algorithm, we run both algorithms over the same RNA sequences. Our algorithm has been implemented in Java programming language and it has been compiled and executed on a computer running Windows 7 as operating system, having 2:4 GHz Intel core i5 processor, and 2 GB of installed memory. To perform the comparison, two standard measures, namely Sensitivity and Specificity, are utilized. These measures are defined as follow:

- Sensitivity: number of correctly predicted base pairs/ number of base pairs in the real structures
- Specificity: number of correctly predicted base pairs/ number of predicted base pairs

A dataset containing nine RNA sequences is employed to perform the comparison. These sequences are randomly selected from [24]. All the structures corresponding to these RNA sequences are native, i.e. they are determined by X-ray crystallography.

The average Sensitivity and Specificity of our genetic algorithm as well as Wiese algorithm are calculated (over 10 independent executions of both algorithms) and presented in Table 1. The values shown in bold face indicate the best results. As it is observed from this table, our proposed extensions greatly improve the accuracy of structure prediction. All the results as well as the implemented source code will be available by request to the corresponding author.

Table 1. Comparing the average Sensitivity and Specificity of our proposed genetic algorithm and Wiese algorithm (over 10 independent executions of both algorithms).

RNA sequence	length	Ours Sensitivity	Ours Specificity	Wiese Sensitivity	Wiese Specificity
ASE_00024	243	**0.7143**	**0.6757**	0.5143	0.5217
ASE_00113	335	**0.6111**	**0.6395**	0.3222	0.3625
ASE_00258	232	**0.7407**	**0.6154**	0.4444	0.3871
ASE_00344	238	**0.7857**	**0.6197**	0.5357	0.4412
PDB_00573	116	0.4118	0.3889	**0.4412**	**0.4545**
PDB_00939	144	**0.1111**	0.098	**0.1111**	**0.1020**
PDB_01144	408	**0.3937**	**0.4386**	0.3592	0.3246
TMR_00272	102	**1.0**	**0.9118**	**1.0**	**0.9118**
CRW_01501	131	**0.8**	**0.8571**	0.5778	0.6341
Average		**0.6187**	**0.5827**	0.4784	0.4599

References

1. Sumitro, Sutiman B. "RECONSIDER OUR UNDERST ANDING ON BIOLOGICAL SYSTEM (A new concept driven by Nanobiology and Complexity Science)." RECONSIDER OUR UNDERST ANDING ON BIOLOGICAL SYSTEM (A new concept driven by Nanobiology and Complexity Science) (2013): 1-38.
2. Major, François, and Richard Griffey. "Computational methods for RNA structure determination." Current opinion in structural biology 11.3 (2001): 282-286.
3. Mathews, David H. "Revolutions in RNA secondary structure prediction." Journal of molecular biology 359.3 (2006): 526-532.
4. Tinoco, Ignacio, and Carlos Bustamante. "How RNA folds." Journal of molecular biology 293.2 (1999): 271-281.
5. Mohammadzadeh, Javad, Mohammad Ganjtabesh, and Abbas Nowzari-Dalini. "Topological properties of RNA variation networks over the space of RNA shapes." MATCH Commun. Math.

Comput. Chem 72 (2014): 501-518.
6. Mohammadzadeh, Javad, Mohammad Ganjtabesh, and Abbas Nowzari-Dalini. "Relation Between RNA Sequences, Structures, and Shapes via Variation Networks." Iranian Journal of Biotechnology 12.3 (2014): 57-70.
7. Zare-Mirakabad, F., et al. "RNAComp: A new method for RNA secondary structure alignment." Match 61.3 (2009): 789.
8. Esmaili-Taheri, Ali, Mohammad Ganjtabesh, and Morteza Mohammad-Noori. "Evolutionary solution for the RNA design problem." Bioinformatics (2014): btu001.
9. Jiang, Tao, Ying Xu, and Michael Q. Zhang. Current topics in computational molecular biology. MIT Press, 2002.
10. Mathews, David H., and Douglas H. Turner. "Prediction of RNA secondary structure by free energy minimization." Current opinion in structural biology 16.3 (2006): 270-278.
11. Nussinov, Ruth, and Ann B. Jacobson. "Fast algorithm for predicting the secondary structure of single-stranded RNA." Proceedings of the National Academy of Sciences 77.11 (1980): 6309-6313.
12. Nussinov, Ruth, and Ignacio Tinoco. "Sequential folding of a messenger RNA molecule." Journal of molecular biology 151.3 (1981): 519-533.
13. Comay, Eliahu, Ruth Nussinov, and Oded Comay. "An accelerated algorithm for calculating the secondary structure of single stranded RNAs." Nucleic acids research 12.1Part1 (1984): 53-66.
14. Hofacker, Ivo L., Martin Fekete, and Peter F. Stadler. "Secondary structure prediction for aligned RNA sequences." Journal of molecular biology 319.5 (2002): 1059-1066.
15. McMellan, N. "RNA Secondary Structure Prediction using Ant Colony Optimisation, Master Thesis." School of Informatics, University of Edinburgh (2006).
16. Wiese, Kay C., and Edward Glen. "A Permutation Based Genetic Algorithm for RNA Secondary Structure Prediction." HIS 87 (2002): 173-182.
17. Zuker, Michael, and Patrick Stiegler. "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information." Nucleic acids research 9.1 (1981): 133-148.
18. Zuker, Michael, and David Sankoff. "RNA secondary structures and their prediction." Bulletin of mathematical biology 46.4 (1984): 591-621.
19. Jacobson, A. B., et al. "Some simple computational methods to improve the folding of large RNAs." Nucleic acids research 12.1Part1 (1984): 45-52.
20. Van Batenburg, F. H. D., Alexander P. Gultyaev, and Cornelis WA Pleij. "An APL-programmed genetic algorithm for the prediction of RNA secondary structure." Journal of Theoretical Biology 174.3 (1995): 269-280.
21. Gultyaev, Alexander P., F. H. D. Van Batenburg, and Cornelis WA Pleij. "The computer simulation of RNA folding pathways using a genetic algorithm." Journal of molecular biology 250.1 (1995): 37-51.
22. Gen, M., and R. Cheng. "Genetic algorithms and Engineering design John Wiley&Sons Inc." 605 Third Avenue (1997): 42-46.
23. Holland, J. H., and D. Goldberg. "Genetic Algorithms in Search, Optimization and Machine Learning." ed: Addison-Wesley, Reading, MA (1989).
24. Andronescu, Mirela, et al. "RNA STRAND: the RNA secondary structure and statistical analysis database." BMC bioinformatics 9.1 (2008): 340.