A genetic algorithm for predicting RNA structures containing pseudoknots

Dongkyu Lee, Kyungsook Han*

School of Computer Science and Engineering, Inha University, Inchon, Korea

*To whom correspondence should be addressed. E-mail: khan@inha.ac.kr

Abstract

This paper describes a genetic algorithm for predicting RNA structures that contain various types of pseudoknots. Pseudoknotted RNA structures are much more difficult to predict by computational methods than RNA secondary structures, as they are more complex and the analysis is time-consuming. We developed an efficient genetic algorithm to predict RNA folding structures containing any type of pseudoknot, as well as a novel initial population method to decrease computational complexity and increase the accuracy of the results. We also used an interaction filter to decrease the size of the possible stem lists for long RNA sequences. We predicted RNA structures using a number of different termination conditions and compared the validity of the results and the times required for the analyses. The algorithm proved able to predict efficiently RNA structures containing various types of pseudoknots in long nucleotide sequences.

Introduction

The prediction of an RNA structure with a pseudoknot using computational methods requires much computation. Predicting the most stable structure with minimal free energy from an RNA sequence is an optimization problem [1, 2, 3]. Computational methods for predicting RNA structure generally make use of two algorithms, one combinatorial the other recursive. The combinatorial algorithm first creates an inventory of all possible stem lists that can be formed by a given RNA sequence, and then determines the combination with the lowest free energy. This algorithm has the advantage that it can include pseudoknot structures, but the number of possible structures increases immensely with sequence length [4, 5]. The recursive algorithm finds the lowest free energy structure from the sub-fragments of a sequence. It makes a systematic search of all sub-fragments for the lowest free energy structure containing at least one base pair. The first sub-fragments considered are those capable of forming a hairpin loop closed by a single base pair. So in a first pass it will find the lowest free energy structures for all pentanucleotides in

the sequence. This method always finds the structure with least free energy, but it does not identify structures such as pseudoknots because of their computational complexity.

A genetic algorithm (GA) is an optimization procedure that implements the mechanism of biological evolution. It begins with a set of solutions called populations. Solutions are then taken and used to form a new population in the hope that the new population will be superior to the old one. They are selected to generate new solutions according to their fitness; the fitter they are, the more opportunities they have to reproduce. This procedure is repeated until some specified condition is satisfied.

GAs have been theoretically and empirically proven to provide robust searches in highly complex and uncertain spaces, and they are finding widespread application in commerce, science and engineering. They are computationally simple and powerful search methods, and many workers have used them to predict RNA structures and sequence alignments; they have been used to seek optimal and sub-optimal secondary RNA structures [6, 7] and to simulate RNA folding pathways [8, 9].

Massively parallel genetic algorithms have been employed to predict RNA structures that include pseudoknots [10, 11]. However the structures predicted contained only H (Hairpin)-type pseudoknots and the computations were extremely complex as they used randomly generated initial populations. Dynamic programming algorithms, also used to predict RNA structures including pseudoknots [4] again could only predict structures with H type pseudoknots, and only from short RNA sequences.

We have developed a GA that is able to predict efficiently RNA structures containing several types of pseudoknots. To predict such structures we derived an approximate energy model for the different types of pseudoknots and developed a topology decision algorithm. To decrease computational complexity, we introduced a long interaction filter and new initial populations methods. We compare and analyze the results predicted by various initial population methods, and also adjust the GA parameters to improve the accuracy of the predictions.

In the section that follows, we describe the GA and outline the new initial population method and the genetic parameters. The implementation of the analysis and the results obtained are given in the following section. Some predicted RNA structures are presented in visual form and their accuracy assessed. General lessons and conclusions are described in the final section.

Prediction algorithm

The prediction algorithm for RNA structures with pseudoknots is composed of two stages: preprocessing, and evolution of the GA. The preprocessing steps reads the RNA sequence and generate three stem pools. From a covariation matrix they generate a list of all possible stems with a minimum of three base pairs. They further calculate the stacking energy of each stem in the stem lists and sort the stems in increasing order of energy values. The list of these stems becomes what we call the *fully zipped stem pool*. Since the number of possible stems increases immensely with sequence length we remove some irregular stems from the stem pools to decrease their size. First, consecutive wobble pairs at either end of a stem are removed because they are not sufficiently stable. The

stacking energy of each stem is then recalculated and irregular stems are removed; these are stems consisting of 1 or 2 base pairs with too distant interactions. These procedures generate the second stem pool that we call the *partially zipped stem pool*. Finally we generate the *pseudoknot stem pool* by finding all possible pairs of stem that can form a typical H type pseudoknot. At this stage, we consider only the number of connecting loops and the length of the pseudoknot stems. The partially zipped stem pools and pseudoknot stem pools are together used to predict RNA structures.

These procedures produce the initial populations that are allowed to evolve using the genetic operator. In using the GA to predict RNA structures, the structures are represented as genome types using binary string genome expression.

Initial population

It is usual to generate initial population randomly when using a genetic algorithm. However this method is not efficient for predicting RNA structures because there are many pairs of stems that cannot coexist in a structure. These stems often share common base pairs or have complex topology, and as a result, randomly generated populations tend to be produce impossible structures. The presence of these impossible structures makes the prediction of RNA structures inefficient. We therefore developed a heuristic method for generating the initial populations.

To generate the initial population, we first select a reference stem in the stem pools, and test the topology of the other stems in stem pools. Topology tests are composed of 2 steps: an overlapping test and crossing test. The overlapping test checks if the stems share base pairs with the reference stem, and the crossing test checks if the stems are crossed with respect to the reference stem and prevents the algorithm from generating complex structures. Every stem in the stem pools is selected as reference stem simultaneously. The complex types of pseudoknots and details of the topology tests are described in the next section.

Two choices have to be made in developing the heuristic initial population. First, as the pseudoknot stem pool is essential when predicting RNA structures containing pseudoknots, the reference stem pool can consist of the pseudoknot stem pool on its own or that pool together with the partially zipped stem pools. The second choice to be made concerns how many stems are included in an RNA structure. One approach is to include all the stems that pass the topology test with the reference stem; the other is to insert only a limited number of these stems in order not to generate complex structures. The number to include is decided in a heuristic manner by repeated testing.

We have compared the results obtained using the four initial population structures derived from combining these alternatives (Table 1). We also generated initial populations by the random method, but the predictions obtained were not good enough to compare with the others.

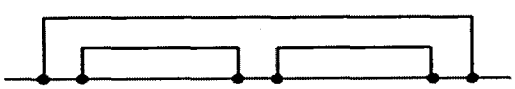
Topology tests

Topology tests are performed to discriminate between the types of loop elements (stems) at the evaluation stage, and to avoid impossible or complex structures at the initial population stage. They are of three types. The overlapping tests and crossing tests are carried

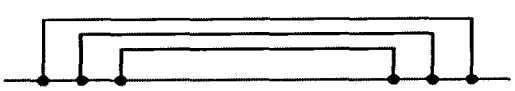
Table 1. Initial population methods

Method	Reference stem pools	Number of stems
1	Partially zipped stem pools with pseudoknot stem pools	no limit
2	Partially zipped stem pools with pseudoknot stem pools	limit
3	Pseudoknot stem pools only	no limit
4	Pseudoknot stem pools only	limit

out at the initial population stage to avoid impossible structures, and the nesting test is performed to decide on the loop types of the stems at the evaluation stage. The latter test checks whether the reference stem contains other nested stems, and is carried out when the reference stem has more than two stems. In effect, it determines whether the topology of the reference stem corresponds to multiple loops, or to nested internal or bulge loops. Figure 1a gives an example of multiple loops and figure 1b of nested loops.



(a) A multiple loop



(b) A nested loop

Figure 1. Example of nested test

Genetic parameters

The performance of GA in solving optimization problems depends on several genetic parameters. These are: the type of genetic operator, the probability of each operator, a fitness function, and the termination conditions.

Crossover and mutation operators are the two basic types of operator. We use a one-point crossover operator that selects one crossover point at random to alter the parental chromosomes, and the crossover is performed with a defined crossover probability. The mutation operator selects bits of the genome at random and inverts them, and mutation is also performed with a defined probability. The crossover operator tends to enable the evolutionary process to move toward promising regions of the search space, and the mutation operator is introduced to prevent premature convergence to local optima; it does so by randomly sampling new points in the search space. We use a high crossover probability and a low mutation probability.

The thermodynamic free energies of RNA structures are used to measure the fitness of individuals in the population. To calculate free energy we use linked list data structures. The node of the linked list is the stem index value of the stem pools. Because the node of the linked list is sorted by order of first index, the loop types of each stem can be easily decided using the topology test. The appropriate energy model is then applied to regular secondary structure elements and H type pseudoknots.

Various types of pseudoknots can be generated during the evolution of the algorithm. In the case of pseudoknots composed of pseudoknot elements and secondary structure elements, thermodynamic free energy can be approximated by current energy models [12]. However for some complex types this is not possible. These complex types are defined in Figure 2, while Figure 3 displays the types of pseudoknots

whose free energy can be calculated, and Figure 4 provides an example of the free energy calculation involved.

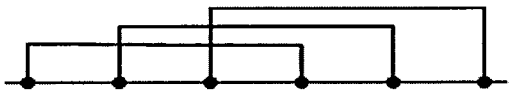


Figure 2. A complex pseudoknot

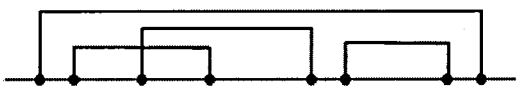


Figure 3. A complex pseudoknot that could be calculated

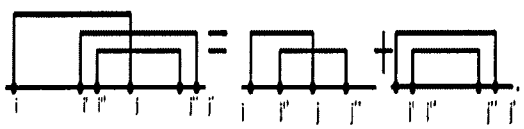


Figure 4. An energy calculation

The termination condition is used to determine whether a genetic algorithm is finished. Since the execution times and accuracy of a prediction algorithm depend on the termination condition, its choice is very important. Either the number of generations, or convergence, can be used as termination condition. Convergence refers to the similarity of the objective scores obtained by comparing the population average score with the score of the best individual in the population: if the population average is within a threshold value, the GA stops evolving. If number of generations is used as termination condition, it is difficult to determine a number that is suitable for all RNA sequences: a relatively low number of generations is required for short RNA sequences, whereas large numbers are generally required for long sequences.

Convergence may be used for all RNA sequences, but the accuracy of prediction is poor because populations tend to converge early. We have predicted structures using both termination conditions and have compared the results.

Results and Discussion

The prediction algorithm was implemented into a program called PseudoFolder with C++ builder 5.0 on a 1.61 GHz Pentium 4 PC with 256 MB memory. PseudoFolder predicts ten structures because the variety of predictions is also important. PseudoFolder integrates the visualization program PseudoViewer [13, 14], so the user can immediately see the predictions in graphical form, and can change the prediction parameters easily using the graphical user interface.

We have used PseudoFolder to predict several structures that include pseudoknots from their sequences. Some of the structures were already known. Figure 5 shows the known structure of TMV RNA and Figure 6 shows the structure of ORSV RNA [15, 16]. These known structure contain either one or two irregular stems as well as irregular pairs that are neither Watson-Crick pairs nor wobble pairs. As these stems could not be included in our predictions because they were not generated during preprocessing, we modified some bases of the test sequences in order to find all the stems represented in the known structures.

We predicted structures five times for each sequence and used average values for statistical analysis, and we compared the accuracy and execution times of the predictions. The accuracy of predictions is defined as the percentage of the known structure elements in the predicted structures, and execution times refer to the time taken by the evolution stage. Figure 7 shows the

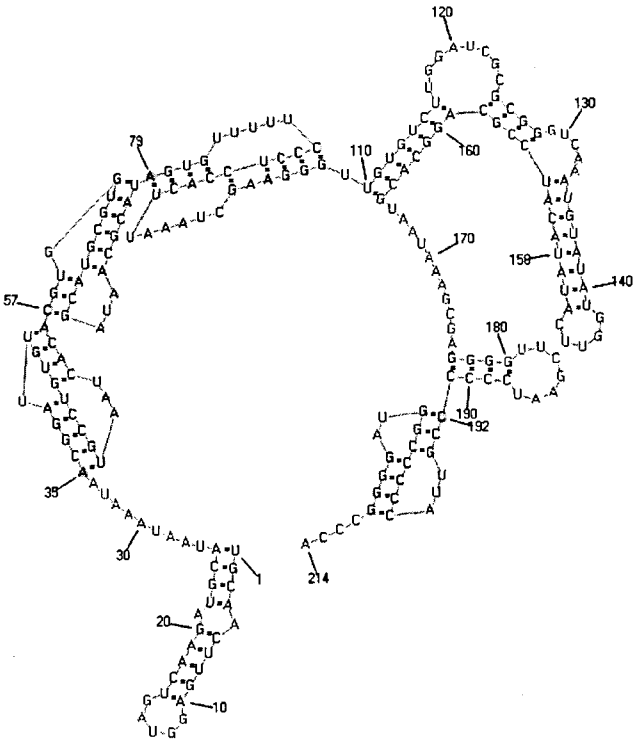


Figure 5. The known structure of TMV

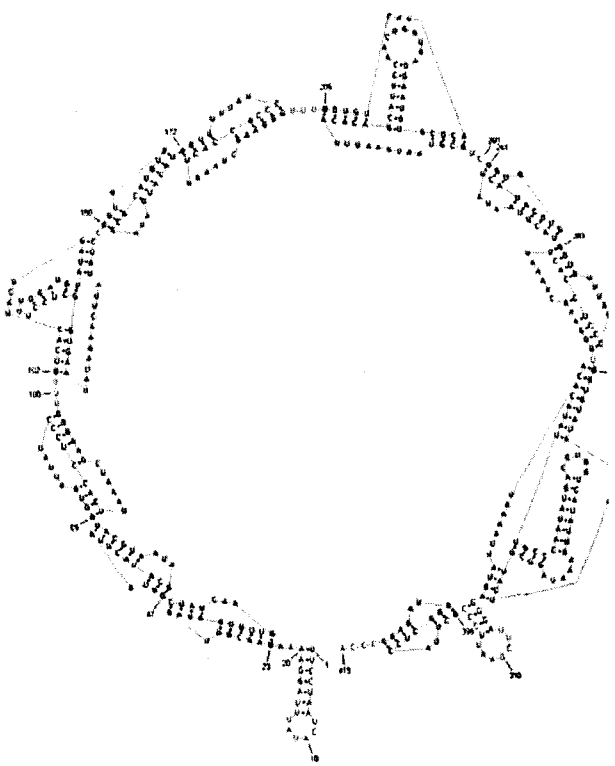


Figure 6. The known structure of ORSV

accuracy of prediction for TMV RNA using convergence as termination condition, and Figure 8 gives the execution times with the same termination condition. The accuracy of prediction was poor. For comparison, Figure 9 and 10 shows the results obtained with number of generations (100) as termination condition. Figure 11 presents the best result obtained for TMV by method 4 with the same termination condition. Thirteen of the predicted stems coincided with those of the known structure and although one stem was different it nevertheless was similar in topology to the known structure. Variation of execution times is small with number of generation as termination condition. And we have been able to decrease execution times when using convergence as termination condition by a method which limits the number of stems.

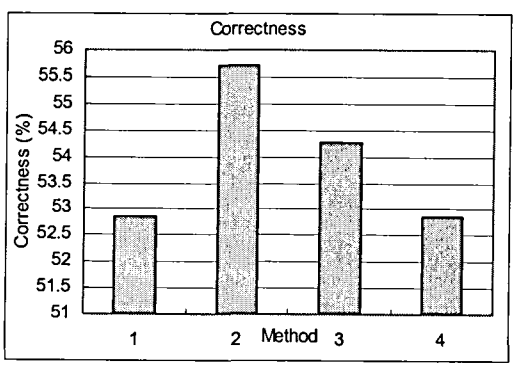


Figure 7. Correctness of prediction results for TMV RNA using convergence as terminal condition

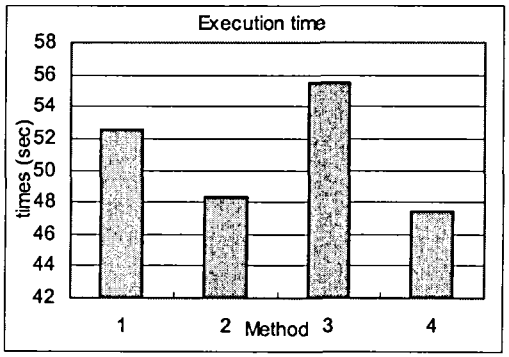


Figure 8. Execution times of predictions for TMV RNA with convergence as termination condition

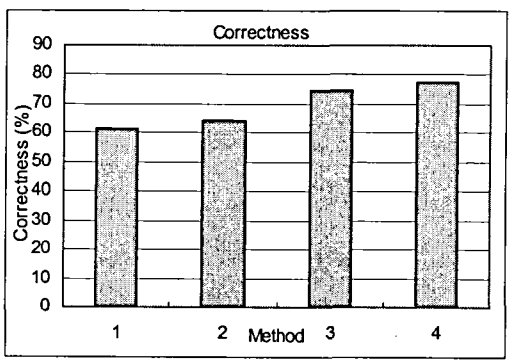


Figure 9. Accuracy of prediction for TMV RNA using number of generations as termination condition (n=100)

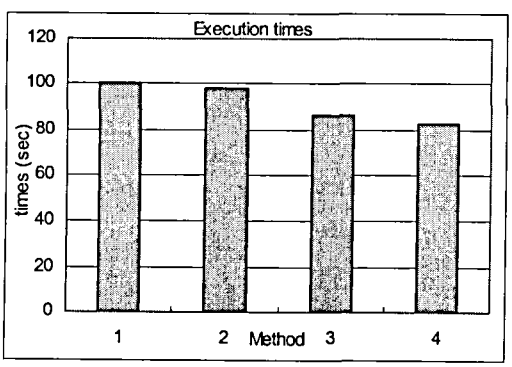


Figure 10. Execution times of predictions for TMV RNA using number of generations as termination condition (n=100)

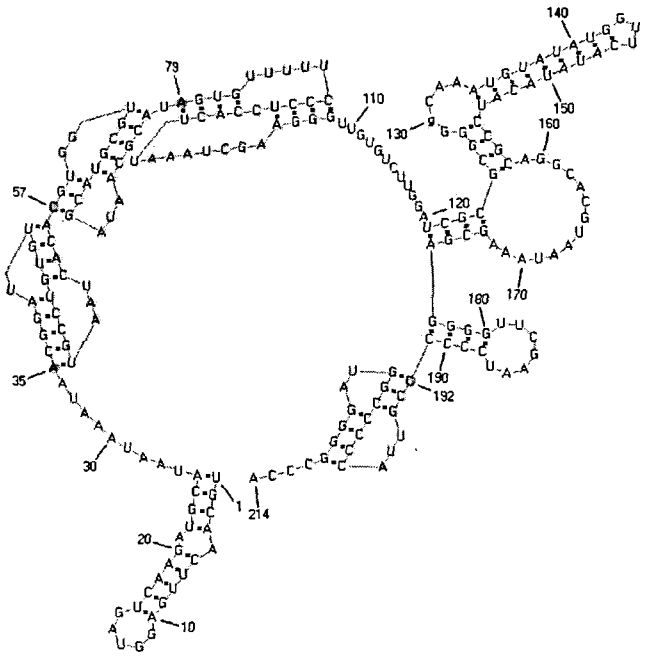


Figure 11. The best prediction for TMV RNA using method 4 and number of generations as termination condition (n=100)

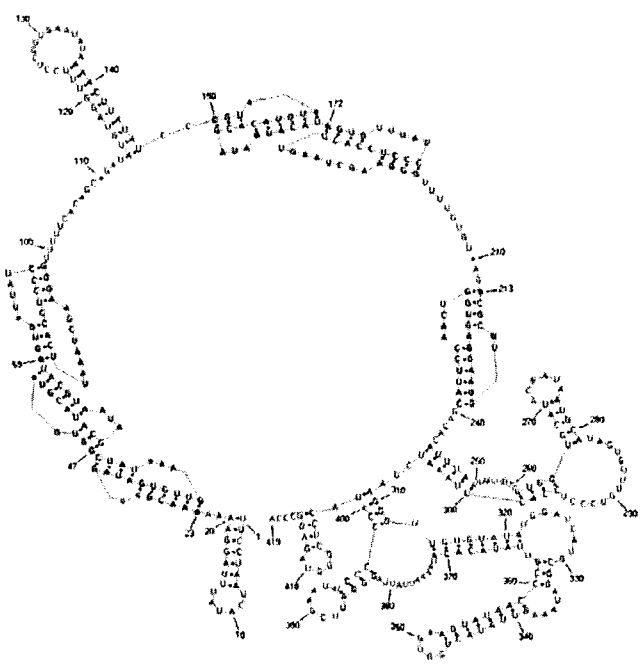


Figure 12. The best prediction for ORSV RNA using method 4 and the number of generations as termination condition (n=300)

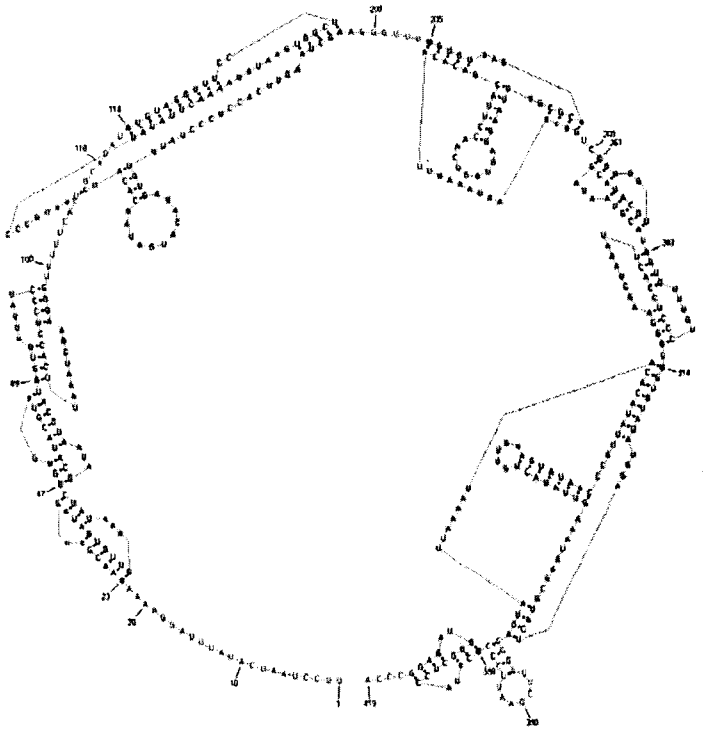


Figure 13. Prediction for ORSV RNA using method 4 that includes various types of pseudoknot

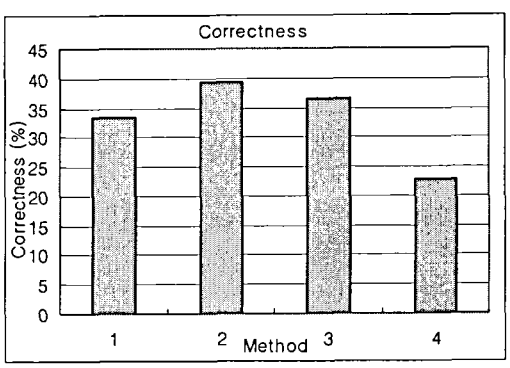


Figure 14. Accuracy of prediction for ORSV RNA using number of generations as termination condition (n=100)

The known structure of ORSV RNA has 28 stems and 8 H type pseudoknots with 3 non classical pseudoknots. The prediction in figure 12 has 18 stems that occur in the known structure and 6 H type pseudoknots. Figure 13 shows another prediction: its accuracy is lower than that of figure 12, but it includes a variety of types of pseudoknots.

In figure 14, we used 100 generations as termination condition. However, the average accuracy of prediction was not satisfactory because ORSV has more bases and stems than TMV RNA. We therefore repeated the prediction with 300 generations as termination condition, and the improved accuracy is shown in figure 15. We also predicted a very similar structure using method 4. Figure 12 shows the best prediction result obtained for ORSV RNA using method 4 and 300 generations as termination condition.

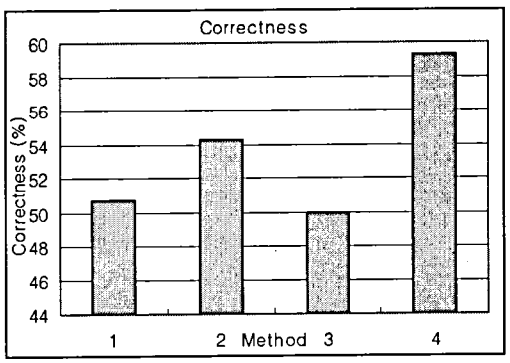


Figure 15. Accuracy of prediction for ORSV RNA using number of generations as termination condition (n=300)

For tests of long RNA sequences, we attempted to predict the optimal structure including pseudoknots using PseudoFolder and a dynamic programming algorithm. However, the dynamic programming algorithm failed to predict the optimal structure because of computational complexity. We cannot therefore guarantee that our prediction algorithm will predict the optimal structure, but it will predict pseudoknot-containing structures similar to the known structure.

The accuracy of predictions depends on many parameters of the genetic algorithm including the initial population method, the control parameter of the genetic operator, the probability of the crossover and mutation operators, and the termination condition. Although it is common practice to use randomly generated initial populations with GA, such randomly generated initial populations proved to be not good enough for RNA structure prediction because there were many stem pairs in the stem pools that could not coexist in a structure. Figure 16 shows the number of stems which overlapped with stems of the known TMV RNA structure. More than 30 stems in the stem pools overlapped with the first stem at the 5' end of the known structure. When we used randomly generated initial populations, the prediction contained overlapping stems and complex structures. In contrast, our initial population method generated simple and stable structure elements capable of evolving.

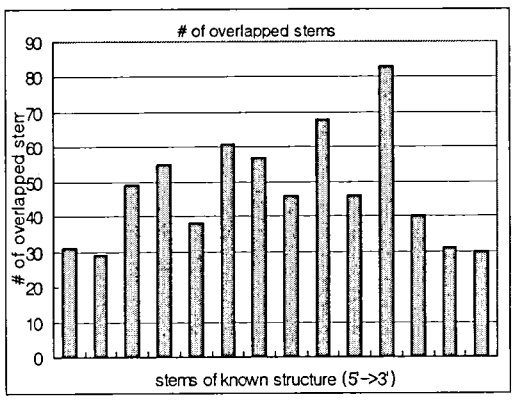


Figure 16. The number of overlapping stems in TMV RNA

Anther important parameter affecting prediction accuracy is the termination condition, which also determined the execution times with the prediction algorithm. For short RNA sequences including simple pseudoknots and secondary structure elements, convergence of population can be used as termination condition, but for long RNA sequence only number of generations is useful for prediction, and the number of generations needed increases with the length of the RNA sequence. As a large number of generations decreases the effectiveness of prediction for short RNA sequences it is difficult to define the number of generations which satisfies accuracy and effectiveness for all RNA sequences.

In view of these considerations we decided to use number of generations as default termination condition to increase the accuracy of the predictions. For short RNA sequences, the user can easily modify the

termination condition and the initial population method using the graphical user interface. The use of new initial population method not only improves the accuracy of prediction but also saves execution times, and we have been able to predict structures with similar topology to known structures using the novel algorithm.

We are currently attempting to develop an algorithm that will decide the optimal control parameter for the GA automatically from the RNA sequence. Although the accuracy of prediction was increased by our approach it still generates many structural elements that differ from the known structure. More refined energy models of the various pseudoknot elements are required to increase the accuracy of prediction We intend to test our prediction algorithm with many more RNA sequences and to improve its performance. We also plan to develop a web-based application program.

Acknowledgements

This work has been supported by the Korea Science and Engineering Foundation (KOSEF) under grant R05-2001-000-01037-0.

References

[1] D. Lee and K. Han, Prediction of RNA pseudoknots-comparative study of genetic algorithm, *Genome Informatics,* 13, 2002, pp. 414-415

[2] D. Lee and K. Han, A Genetic Algorithm for Predicting RNA Pseudoknot Structures, *Lecture Notes in Artificial Intelligence,* 2843, 2003, pp. 333-340

[3] B. A. Deiman and C. W. A. Pleij, A vital feature in viral RNA, *Seminars in Virology,* 8, 1997, pp. 166-175

[4] E. Rivas and S. R. Eddy, A dynamic programming algorithm for RNA structure prediction including pseudoknots, *Journal of Molecular Biology,* 285, 1999, pp. 2058-2068

[5] T. Akutsu, Dynamic programming algorithm for RNA secondary structure prediction with pseudoknots, *Discrete Applied Mathematics,* 104, 2000, pp. 45-62

[6] G. Benedetti, and S. Morosetti, A genetic algorithm to search for optimal and suboptimal RNA secondary structure, *Biophysical Chemistry,* 55, 1995, pp. 253-259

[7] B. A. Shapiro and J. Navetta, A massively parallel genetic algorithms for RNA secondary structure prediction, *Journal of supercomputing,* 8, 1994, pp. 195-207

[8] A. P. Gultyaev, F. H. D. van Batenburg and C. W. A. Pleij, The computer simulation of RNA folding pathway using a genetic algorithm, *Journal of Molecular Biology,* 250, 1995, pp. 37-51

[9] B. A. Shapiro, J. C. Wu, D. Bengali and M. J. Potts, The massively parallel genetic algorithms for RNA folding: MIMD implementation and population variation, *Bioinformatics,* 17, 2001, pp. 137-148

[10] B. A. Shapiro and J. C. Wu, An annealing mutation operator in the genetic algorithms for RNA folding, *Computer Applications in the Biosciences,* 12, 1996, pp. 171-180

[11] B. A. Shapiro and J. C. Wu, Predicting RNA H-type pseudoknots with the massively parallel genetic algorithm, *Computer Applications in the Biosciences,* 13, 1997, pp. 459-471

[12] J. P. Abrahams, M. van den Berg, E. van Batenburg and C. Pleij, Prediction of

RNA secondary structure, including pseudoknotting, by computer simulation, *Nucliec Acids Research,* 18, 1990, pp. 3035-3044

[13] K. Han, Y. Lee and W. Kim, PseudoViewer: automatic visualization of RNA pseudoknots, *Bioinformatics,* 18, 2002, pp. S321-S328

[14] K. Han and Y. Byub, PseudoViewer2: visualization of RNA pseudoknot of any type, *Nucleic Acids Res,* 28, 2003, pp. 3432-3440

[15] C. Einvik, H. Nielsen, R. Nour and S. Johansen, Flanking sequences with an essential role in hydrolysis of a self-cleaving group, I-like ribozyme, *Nucliec Acids Res,* 28, 2000, pp. 2194-2200

[16] A. P. Gultyaev, E. van Batenburg and C. W. A. Pleij, Similarities between the secondary structure of satellite tobacco mosaic virus and tobamovirus RNAs, *Journal of General Virology,* 75, 1994, pp. 2851-2856