A Genetic Programming Based System for the Prediction of Secondary and Tertiary Structures of RNA

Katsuhisa Yamaguchi [1] **Carlos Adriel Del Carpio** [2]

yamaguti@translell.eco.tut.ac.jp carlos@translell.eco.tut.ac.jp

[1] Laboratory for Informatics & AI in Molecular and Biological Sciences, Department of Knowledge-based Information Engineering, Toyohashi University of Technology, Tempaku, Toyohashi 441-8580, Japan

[2] Laboratory for Informatics & AI in Molecular and Biological Sciences, Department of Ecological Engineering, Toyohashi University of Technology, Tempaku, Toyohashi 441-8580, Japan

1 Introduction

Several attempts to predict automatically the RNA secondary structure have been performed in recent years[1,2]. These attempts can be divided in essentially two general approaches. The first involves the overall free energy minimization by adding contributions from each base pair, bulged base, loop and other elements[1]. The second type of approach is more empirical and involves searching for the combination of non-exclusive helices with a maximum number of base pairing [2].

Within the latter, methods using DP (dynamic programming) are the most common [2,3]. Here we introduce a new concept in prediction of RNA structures, and we extend the hitherto existing secondary structure prediction systems into the next step i.e. the prediction of the tertiary structure of the macromolecule from the predicted secondary structure. This will allow the identification of receptor regions on the molecule as well as detailed evaluation of its biochemical an biological functions.

2 Methodology

Stated in the simplest manner, the problem of predicting the secondary structure of RNA structures consists in maximizing the number of base pairings (GU,GC,AU), satisfying the condition of a tree like structure for the bio-molecule. A DP recurrent procedure solves the problem in $O(n^3)$ time [4], however consideration of further elements increases the time to $O(n^4)$ and $O(n^5)$ [4]. Furthermore, the prediction ability is still far below optimal.

In the present work we introduce a heuristic algorithm based on the genetic programming paradigm (GP) [5]. The problem is encoded as a parsing tree of possible configurations of the molecule. Each tree is a chromosome representing a configuration for the molecule. The three operations of GP, i.e. selection, crossover,and mutation are applied to a population of chromosomes or individuals.

The evolution process has as penalty function the number of bases pairings, and the process evolves so as to maximize this number. Each chromosome or individual records explicitly the number of pairings for a determined segment of the molecule.

At the end of the evolution process, the program yields a set of chromosomes with high scores (number of bases pairings), containing with high probability structures analogous to the native RNA structure.

The process, however, does not end with the prediction of the secondary structure, but goes further into attempting the prediction of the three dimensional structure of the biomolecule. To carry out this, the information for the secondary structure is transferred to a distance matrix. As the RNA molecules contain hundreds of bases, and each base contains several atoms, handling the entire distance matrix would be too costly. Instead we have developed a partial DG (Distance Geometry) methodology, which involves the prediction of the 3D structure of the molecule by regions where pairing is high. Thus, the 3D structure is initially built also by 3D segments. Connection of the so generated segments into a single molecule is carried out by minimization of the internal force field of the entire molecule.

3 Results and Discussion

Here we present some preliminary results of the system described above. In Fig. 1, we present one of the individuals of the last generation of evolved structures. The structure is represented by the corresponding parse tree. Fig. 2 is the secondary structure of a small fragment of a RNA molecule.

Although we have not yet succeeded in predicting the structure of real RNA structures, the prediction of the secondary structure of medium sized RNA sequences makes of the method a promising one. The introduction of the new concept of evolution into the prediction of RNA structure is thus relevant. Finally, prediction of the 3D structure of the macromolecule is also of the utmost importance, to elucidate functional characteristics such as protein binding regions, activity areas, and to give a deeper insight into protein synthesis in living systems.

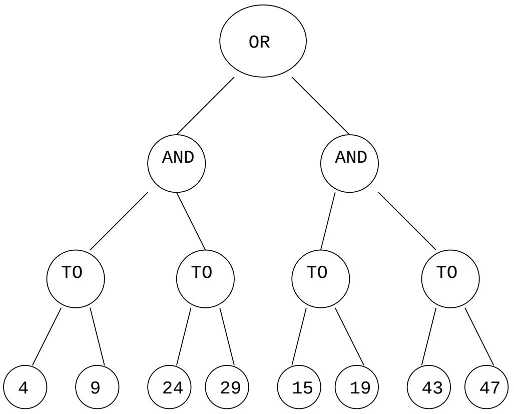

Figure 1: A parse tree representing the configuration of the RNA molecule.

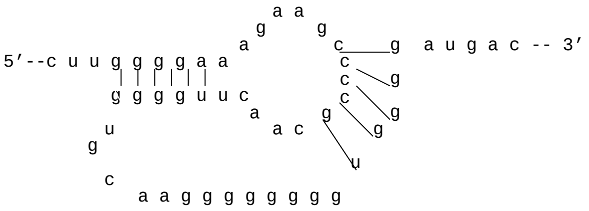

Figure 2: Configuration of the RNA dictated by the chromosome in Fig. 1.

References

[1] Adrahams, J.P. and Breg, M., Prediction of RNA secondary structure including pseudoknotting by computer simulation, *Nucleic Acids Research*, 18:3035–3044, 1990.

[2] Waterman, M., RNA structure prediction, Methods in Enzymology, Academic Press, San Diego, Vol. 164, 1988.

[3] Zuker, M. and Stiegler, P., Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, *Nucleic Acids Research*, 9:133–148, 1981.

[4] Akutsu T., DP Algorithms for RNA secondary structures Prediction with Pseudoknots, *Genome Informatics 1997*, Universal Academy Press, 8:173–179, 1997.

[5] Koza, J.R., *Genetic Programming*, MIT Press, 1992.