Evolving Finite Automata with Two-Dimensional Output for DNA Recognition and Visualization

Edgar E. Vallejo
Computer Science Department
Tecnológico de Monterrey
Campus Estado de México, México
evallejo@campus.cem.itesm.mx

Fernando Ramos
Computer Science Department
Tecnológico de Monterrey
Campus Morelos, México
framos@campus.mor.itesm.mx

Abstract

This article presents a computational model for biosequence recognition and visualization. Finite automata with 2-D output recognize and create maps of DNA sequences. Populations of such machines evolve discrimination capabilities using genetic algorithms. Experimental results indicate that evolved machines are capable of recognizing HIV sequences in a collection of training and validation sets. In addition, we found that these sequences map to similar structures on the cartesian space.

1 FINITE AUTOMATA WITH TWO-DIMENSIONAL OUTPUT

A finite automaton with two-dimensional output is a special type of finite automata (Hopcroft and Ullman, 1979). It has a read-only input tape. The input tape head can read one symbol at a time, and in one move the input tape head shifts one square to the right. The output of the machine is produced on the cartesian space. The output tape head writes one symbol at the current position and in one move the output tape head advances one coordinate (up, down, left or right). The machine either accept or reject each sequence.

2 EXPERIMENTS

We applied a genetic algorithm to a population of finite automata with two-dimensional output to evolve the ability of discriminating a collection of DNA sequences from a collection of training and validation sets. These sets consist of positive and negative examples of HIV sequences. We used a generational genetic algorithm with tournament selection and elitism (Mitchell, 1996). Genome representation was derived from the concatenation of values of the finite automata state transition function (Vallejo and Ramos, 2001). Fitness was defined as the number of sequences correctly classified.

3 RESULTS

We performed several runs with this model. In most runs, the evolutionary process yielded finite automata capable of classifying all sequences correctly. In addition, we found that HIV sequences map to very similar structures on the cartesian space.

4 CONCLUSIONS

In this work, we extended the finite automata model to perform biosequence recognition and two-dimensional output. We evolved a population of such machines to perform sequence recognition. The evolutionary process yielded machines with recognition and generalization capabilities. The evolved machines were capable of producing a convenient biosequence representation.

References

J.E. Hopcroft, J.D. Ullman (1979). *Introduction to Automata Theory, Languages and Computation*. Addison Wesley Publishing Company.

M. Mitchell (1996). *An Introduction to Genetic Algorithms*. The MIT Press.

D.W. Mount (2001). *Bioinformatics: Sequence and Genome Analysis*. Cold Sring Harbor Laboratory.

E.E Vallejo, F. Ramos (2001). Evolving Turing Machines for Biosequence Recognition and Analysis. In J. Miller, M. Tomassini, P.L. Lanzi, C. Ryan, A.G.B. Tettamanzi W.B. Langdon (eds.) *4th European Conference on Genetic Programming, EuroGP2001* LNCS 2038. pp. 192-203. Springer-Verlag.