DNA Sequence Analysis using Hidden Markov Model and Genetic Algorithm

Tetsushi Yada [1] Masato Ishikawa [2] Hidetoshi Tanaka [2] Kiyoshi Asai [3]
yada@mri.co.jp ishikawa@icot.or.jp htanaka@icot.or.jp asai@etl.go.jp

[1] Systems Science Department, Mitsubishi Research Institute, Inc.
[2] Institute for New Generation Computer Technology (ICOT)
[3] Electrotechnical Laboratory (ETL)

We have developed a new method for the analysis of DNA sequences. The method consists of the Hidden Markov Model (HMM) [1] and Genetic Algorithm (GA) [2], and can identify conserved patterns, can estimate the distances between the patterns and can classify the sequences. We have applied the developed method to the extraction of conserved patterns in the vicinity of the 5' end of yeast intron DNA sequences, and have obtained preliminary results.

HMM is a kind of stochastic model that is defined as a nondeterministic finite state automaton represented by a Markov process. Since it has been shown that HMM is capable of representing the characteristic features in DNA and protein sequences [3], there are some activities for analyzing the sequences using HMM. In such an analysis, obtaining an optimal topology of the HMM network for given sequences is an important problem. On the other hand, our current knowledge of the sequences is not enough for the efficient design of the topology. In this case, the problem can be regarded as a kind of combinational problem with discrete variables in which a "combinatorial explosion" frequently occurs. Although some methods [4, 5] have been proposed for the problem, there are some problems with computational time and constraints of the topology growth.

The main purpose of the present study is to develop a generating method for preferable HMM network topology. Therefore, to increase the efficiency of the topology generation, we have adopted GA. It is a kind of stochastic search method developed originally in a field of mathematical biology and simulates the evolution of a population.

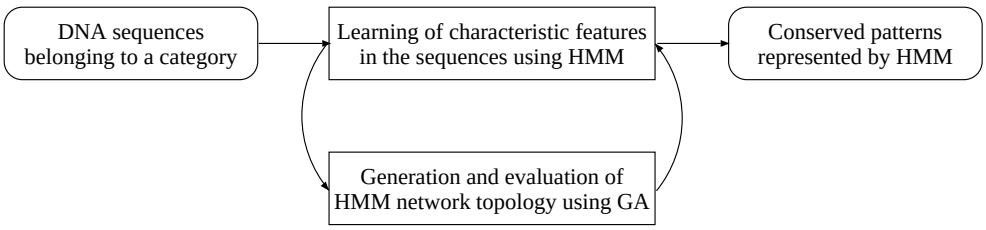

Figure 1: Schematic diagram of the method

A schematic diagram of the method is shown in Figure 1. Input data is a set of DNA sequences which belongs to a category. The method searches an optimal topology of the HMM network for the data set, while repeating the efficient generation and evaluation of the topology. At the begining of a search, HMMs are generated randomly, and each of them evolves according to the manner of GA. In general, the likelihood of HMM increases with the complexity of the topology. However, it is known that over representation is frequently observed as the complexity increases [6]. Therefore, in order to balance the likelihood and the complexity, we have adopted artificial fitness based on Akaike Information Criterion (AIC) [7]. The fitness W_i of the ith HMM in an artificial HMM population is given by the following equation:

$$W_i = [-2 \log L_i(\theta; f) + 2\lambda p_i]^{-1} \quad (1)$$

[1]矢田 哲士: (株) 三菱総合研究所 システム科学部, 〒 100 東京都千代田区大手町 2-3-6
[2]石川 幹人, 田中 秀俊: (財) 新世代コンピュータ技術開発機構 第 2 研究部, 〒 108 東京都港区三田 1-4-28
[3]浅井 潔: 電子技術総合研究所 知能情報部, 〒 305 つくば市梅園 1-1-4

where $\log L_i(\theta; f)$ is a maximum logarithm likelihood of the ith HMM which is obtained using the Baum-Welch algorithm, p_i is the number of free parameters which exist in the ith HMM, and λ is the balancing factor.

The topology is encoded into an artificial chromosome using the strong specification method [8], and the growth and contraction of the topology occur by artificial insert and delete mutations, respectively. In order to avoid converging in the local maximum, we have adopted coding techniques [9] and models [10] based on population genetics. Moreover, parameters specifying an HMM, which are initial values of state transition probabilities and output symbol distribution at each state, are encoded into artificial chromosomes, as well as the topology.

We have applied the newly developed method to the extraction of conserved patterns in the vicinity of the 5' end of yeast intron DNA sequences [11]. Each sequence consists of 25 base pairs. The artificial population at the initial state of a search consists of two state HMMs generated randomly. We have performed computer experiments under the following conditions. (1) the population size is 50. (2) the number of search generations is 100. (3) the balancing factor λ is 1.0. As a result, we have obtained an HMM with the maximum artificial fitness, which consists of six states. The HMM is shown in Figure 2, and implies that a GTATG pattern is a consensus sequence of the 5' end of yeast introns, and AT-rich sequences locate downstream of the pattern. GT at the 5' end of the pattern corresponds to the GT dinucleotide of the GT-AG rule. The results indicate that the method is capable of the efficient generation of HMM network topology, and the extraction of conserved patterns in DNA sequences.

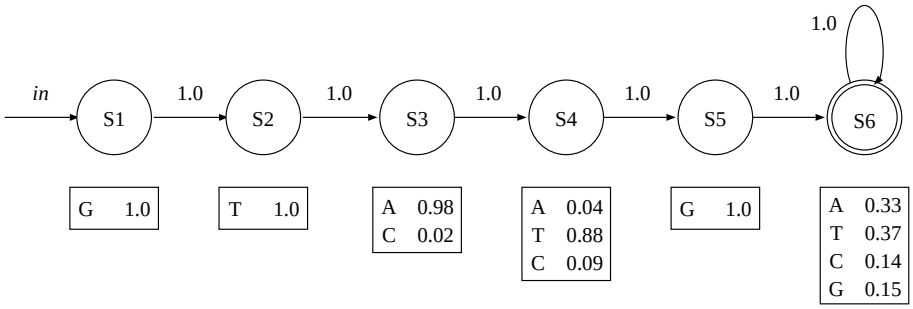

Figure 2: An HMM obtained using the method

Our plans for future study are to apply the newly developed method to the analysis of primate DNA promoter sequences, and to further investigate the efficiency of the method.

References

[1] Levinson, S. E. et al.: *Bell Syst. Tech. J.*, **62**, 1035-1074 1983.

[2] Holland, J.: *Adaptation in Natural and Artificial Systems*, MIT Press, 1992.

[3] Asai, K. et al.: In *Proceedings Genome Informatics Workshop II*, 144-147 1991.

[4] Tanaka, H. et al.: In *Proceedings Genome Informatics Workshop IV*, 224-230 1993.

[5] Fujiwara, Y. and Konagaya, A.: In *Proceedings Genome Informatics Workshop IV*, 56-64 1993.

[6] Konagaya,A. and Kondo,Y.: In *Hawaii Int. Conf. on System Sciences*, 746-755 1993.

[7] Akaike, H: In *Proceedings of the 2nd Int. Symp. on Information Theory*, 267-281 1973.

[8] Miller, G. et al.: *Proc. of ICGA-89*, 1989.

[9] Goldberg, D.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[10] Hartl, D. L. and Clark, A. G.: *Principles of Population Genetics*, Sinauer, 1989.

[11] Konopka, A. K.: In *Proceedings of the 2nd International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, 69-87 1993.