Extraction of Hidden Markov Model Representations of Signal Patterns in DNA Sequences

Tetsushi Yada

The Japan Information Center of Science and Technology (JICST)
5-3 Yonbancho, Chiyoda-ku, Tokyo 102, Japan
Phone: +81-3-5214-8491
yada@jicst.go.jp

Masato Ishikawa†, Hidetoshi Tanaka‡ and Kiyoshi Asai*

†*Matsushita Electric Industrial Co.,Ltd.*
ishikawa@trl.mei.co.jp

‡*Mitsubishi Electric Corp.*
htanaka@isl.melco.co.jp

* *Electrotechnical Laboratory (ETL)*
asai@etl.go.jp

Keywords: *DNA sequence; signal pattern extraction and recognition; Hidden Markov Model; Genetic Algorithm; Akaike Information Criterion*

Abstract

We have developed a method to extract the signal patterns in DNA sequences. In this method, the Genetic Algorithm (GA) and Baum-Welch algorithm are used to obtain the best Hidden Markov Model (HMM) representations of the signal patterns in DNA sequences. The GA is used to search the best network shapes and the initial parameters of the HMMs. Baum-Welch algorithm is used to optimize the HMM parameters for the given network shapes. Akaike Information Criterion (AIC), which gives a criterion for the balance of adaptation and complexity of a model, is applied in the HMM evaluation. We have applied the method to the extraction of the signal patterns in human promoters and 5' ends of yeast introns. As a result, we obtained HMM representations of characteristic features in these sequences. To validate the efficiency of the method, we have performed promoter recognition using obtained HMMs. Two entries including nine promoters are selected from GenBank 76.0, and it is observed that the HMM can predicts eight promoters correctly. These results imply that the method is efficient to design preferable HMM networks, and provides reliable models for the recognition of the signal patterns.

1 Introduction

As the advance in DNA sequencing projects, enormous DNA sequences have been stored. In order to understand the meaning of the sequences, it is important to clarify the mechanism of expression and control of the genetic information in DNA sequences. It is known that the signals in DNA sequences contain biologically important information about the function and the evolution. However, mutations accumulated in sequences yielded the diversities of the sequences. The sequences in functional regions possess following diversities: the diversity of (1) signal sequences, (2) distances between signals, and (3) permutation and combination of signals. For instance, it is known that promoters contain several kinds of signals having various compositions and locations [Bucher 90]. Further, permutations and combinations of the signals are different in promoters [Levin 94]. The signal patterns are usually expressed by regular patterns, but it is difficult to capture these diversities by such simple regular patterns. Reliable methods to extract the flexible signal patterns from the data has been desired.

Stochastic models are useful for the representation of the diversity of the signal patterns. In this study, we have adopted Hidden Markov Models (HMMs) [Levinson *et al.* 83] as the representations of the signal patterns. An HMM is defined as a nondeterministic finite state automaton represented by a Markov process. It has been shown that HMMs are capable to represent the characteristic features in the sequences [Asai *et al.* 93, Haussler *et al.* 93].

We used HMMs whose output symbols of the 'hidden' states are 'A', 'T', 'G', 'C'. The parameters of an HMM, the transition probabilities and the output distribution, can be trained using Baum-Welch Algorithm to maximize the likelihood of the HMM to the DNA sequences. However, these algorithms require that the network shape of the HMM is fixed before the learning. Therefore, it is important to obtain the optimal HMM network for the given sequence. Since it is difficult to know which HMM network is optimal before the training of the HMM, we have to repeat network design and parameter training by trial and error. Several methods, [Fujiwara *et al.* 94, Tanaka *et al.* 93] have been proposed for designing HMM networks. However, these methods have problems with constraints of the HMM topology and the dependency of the initial parameter values.

The main purpose of this study is to develop an efficient method to generate the optimal HMM networks and to get the optimal HMM representations of the signal patterns of the DNA sequences. We have developed the method which uses the Genetic Algorithm (GA) [Holland 92] to design the HMM network and initial parameters of the HMM. The HMM networks and the initial parameters are encoded into artificial chromosomes, and the GA is operated with these chromosomes to find the optimal HMM network and the initial parameters. The fitness of an chromosomes is calculated by training the HMM, whose network and initial parameters are represented by the chromosomes, by Baum-Welch Algorithm with the DNA sequences whose signal patterns we want to model.

2 Method

We have developed the method which uses the Genetic Algorithm (GA) to design the HMM network and initial parameters of the HMM. The GA is a heuristic algorithm to search the optimal solutions of problems, which simulates the biological evolution process. Each individual has chromosomes, where a possible solution of the problem is encoded. The individuals of the population of new generation are produced by mating the individuals which have high fitness in the population of previous generation. The fitness is the performance of the solution which is encoded by the chromosomes of the individual. During the process of mating, recombinations and mutations occur by some probabilities. These recombinations and mutations have the effect of generating new types of solutions and of avoiding local optima.

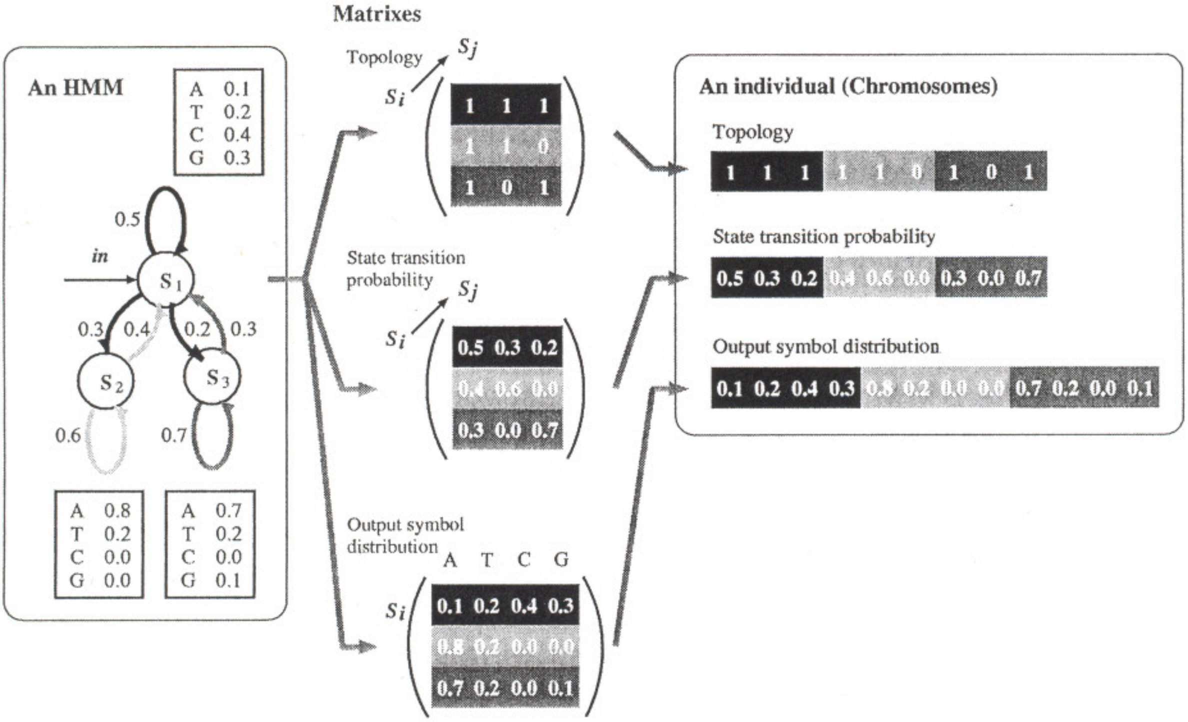


Figure 1: Encoding of an HMM network and parameters

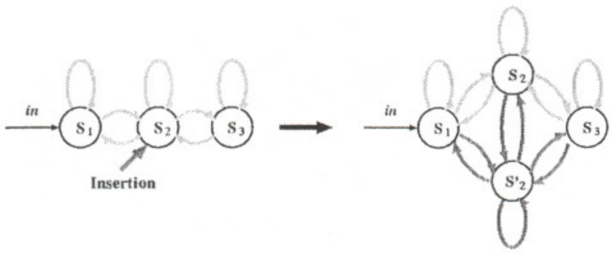


Figure 2: Insertion in HMM network

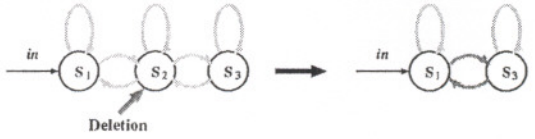


Figure 3: Deletion in HMM network

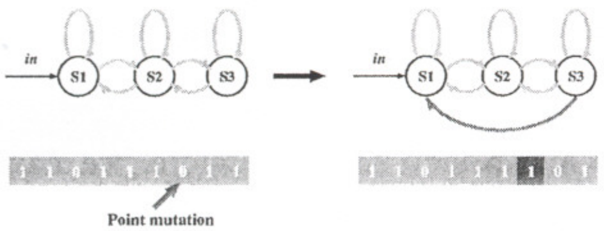


Figure 4: Point mutation in HMM network

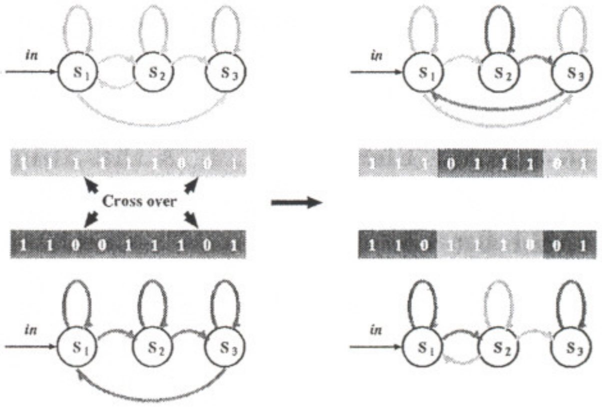


Figure 5: Recombination in HMM network

In the method, the HMM network and the initial parameters are encoded into 3 chromosomes of each individual. The first chromosome encodes the HMM network, the second chromosome encodes the initial transition probabilities, and the third chromosome encodes the initial output distributions. Figure 1 illustrates the encoding of an HMM network and parameters into an chromosome.

The method searches the optimal solution in the space of any kind of HMM network, because the encoding uses connection matrix to represent the network topology. Figure 2∼5 illustrate the mutations and the recombination, which are performed on the HMM networks during the mating process. We have designed insert mutation so that the topology after the mutation always includes the original one. With the help of this operation, the method can perform the accumulative development of the topology.

The probability that an individual is used for mating is defined proportional to the fitness of the individuals. The fitness of each individual is calculated as follows. The HMM, whose network and initial parameters are encoded by the chromosomes of the individual, is trained by Baum-Welch algorithm using the given DNA sequences. This training optimize the parameters of HMM to fit the sequences. By using Akaike Information Criterion (AIC)[Akaike 73], the fitness W_i of the ith individual (HMM) in the population is given by the following equations:

$$W_i = w_i / \sum_{j=1}^{N} w_j \tag{1}$$

with

$$w_i = \left[-2\log L(\hat{\theta}_i; f) + 2\lambda p_i\right]^{-1} \tag{2}$$

N is the number of HMMs in the population. $\hat{\theta}_i$ and f indicate the ith HMM parameters optimized by Baum-Welch algorithm and the set of DNA sequences. $\log L(\hat{\theta}_i; f)$ is the maximum logarithm likelihood estimates of the ith HMM. The p_i, which is the number of free parameters consisting of the ith HMM, indicates the complexity of the HMM. In general, the likelihood of HMM for the given sequences increases with the complexity of the network increases. However, it is known that over representation is frequently observed as the complexity increases [Konagaya and Kondo 93]. Therefore, we have considered the balance of the likelihood and the complexity.

3 Data

According to the procedures described below, we have created three data sets of human DNA sequences and one data set of yeast DNA sequences, CAAT boxes, TATA boxes, promoters and 5' ends of introns.

From GenBank release 76.0 [GenBank 93], we selected three groups of human entries. The entries of the first group have clear descriptions of CAAT box in the feature tables, and we took sequences in the vicinity of CAAT box from the entries. In order to remove bias from the sequences, we selected the sequences with less than 70 % similarity. The selected sequences are regarded as a CAAT box data set. The second group consists of the entries having clear descriptions of TATA box in the feature tables. From the second group, we created a TATA box data set by applying the procedures similar to the CAAT box data set. Each sequence in both sets consists of 36 bases. The third group consists of the entries having clear descriptions of both CAAT box and TATA box in the feature tables. From the third group, we created a promoter data set. Each sequence in the set consists of 71 bases. From NRFES (Non-Redundant Functionally Equivalent Sequences) version 0.5 [Konopka 93], we collected DNA sequences in the vicinity of 5' ends of introns. Each sequence consists of 25 bp.

By applying the procedures described above, we have collected 168, 123, 55 and 57 sequences as the CAAT box, TATA box, promoter data set and the 5' ends of introns, respectively.

Table 1: Parameters of the method

Population size	50
Searching generation	100
Recombination rate	0.10
Point mutation rate	0.05
Insert mutation rate	0.04
Delete mutation rate	0.02
Number of states at first generation	2
Balancing factor	0.10
Maximum iteration number of Baum-Welch algorithm	50

4 Results

We have applied the method for the extractions of the signal patterns from four data sets described above. Parameters of the method are shown in Table 1. Figure 6∼9 show HMMs obtained by the method using CAAT box, TATA box, promoter data set and 5' ends of introns, respectively. The filled states, the solid output symbol distributions and the solid state transitions correspond to the signal patterns which are extracted from the data sets. In this study, we have defined that state transition starts at the first number of state and ends at the last number of state for all sequences in a data set.

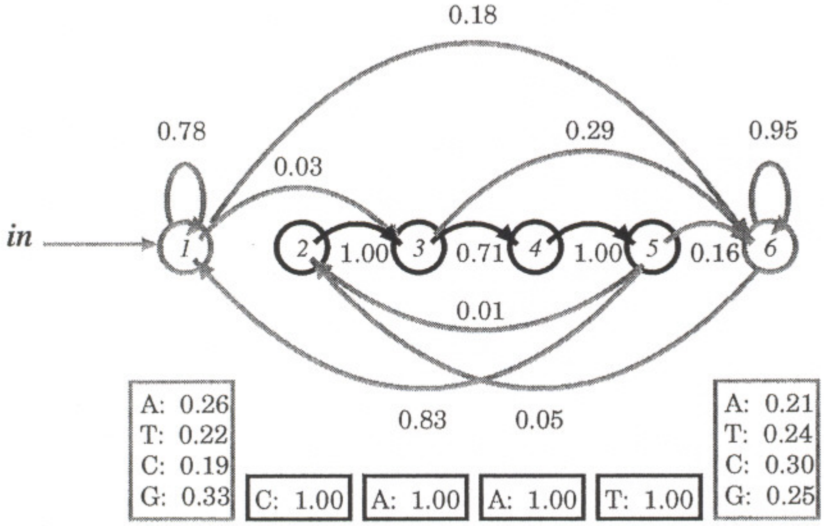


Figure 6: A CAAT box HMM obtained by the method

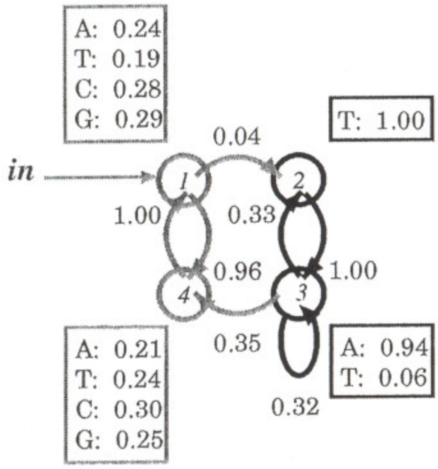


Figure 7: A TATA box HMM obtained by the method

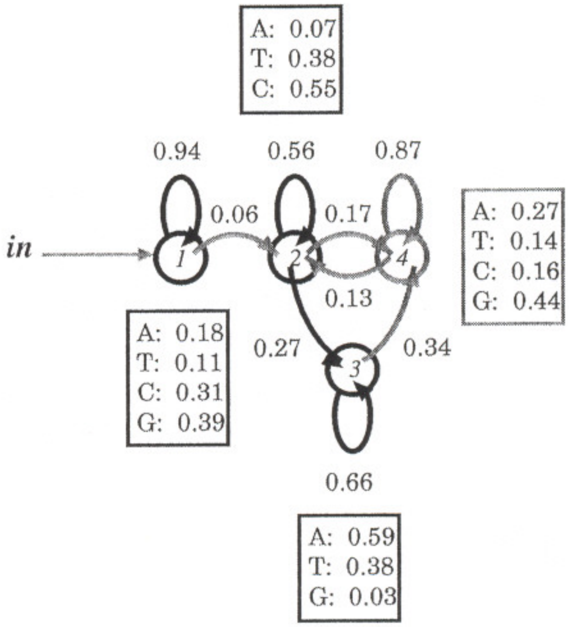


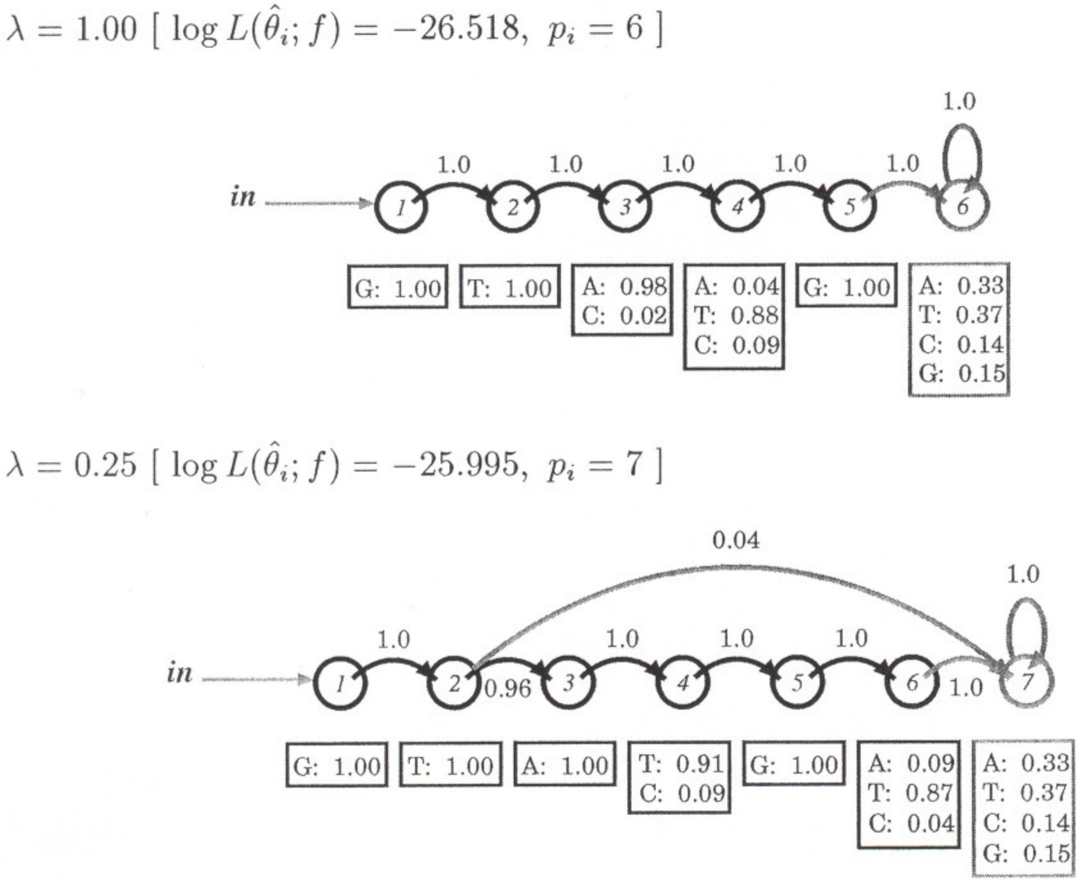
Figure 8: A promoter region HMM obtained by the method

In Figure 6, CAAT box is represented by state transition *2* → *3* → *4* → *5*, and it is obvious that "CAAT" subsequence can be regarded as consensus pattern of CAAT box. In Figure 7, TATA box is represented by state transition between states *2* and *3*. The transition represents TATA box as AT-rich subsequence which starts with base T. State *1* and *6* in Figure 6 and state *1* and *4* in Figure 7 are dummy states which indicate that no significant bias of base composition is observed in the vicinity of either box. In Figure 8, the HMM represents three kinds of signal sequences. CAAT and TATA box are represented by states *2* and *3*. State *1* represents GC box of GC-rich subsequence located upstream of CAAT and TATA box. Note that we were not conscious of GC box in collecting promoter sequences. However, the HMM dose not specify the positional relation between CAAT and TATA box. Further search is required for more detailed representations.

Figure 9 shows the effect of balancing factor λ in Equation 2. Both HMMs represents that AT-rich sequences are located downstream from the pattern. However, the upper HMM represents "GTATG" pattern as a consensus sequence and the lower HMM represents "GTATGT" pattern as a seconsensus sequence. The upper HMM has the simpler network than the lower HMM, but maximum likelihood of upper one is smaller than that of lower one. That indicates the factor λ has the capability to design the balance of likelihood and complexity of the HMMs.

In order to validate the efficiency of the method, we have performed promoter recognition using the HMM shown in Figure 8. From GenBank 76.0, we have selected two entries including nine promoters. Figure 10 shows the result of the recognition. The horizontal axis indicates the position in the DNA sequence. The vertical axis indicates the score of DNA subsequence. Filled bars show that the HMM recognizes subsequences as promoter regions. Arrows shows the positions of promoters. It is observed that the HMM can predict eight promoters correctly, but several false positives are observed. The HMM contains characteristic features in human promoter region and it is surprising that the features of complex promoter are encoded into only 23 parameters in the HMM.

As seen from Figures 6~9 and Figure 10, the method is capable of extracting signal patterns from given sequences and provides a reliable model for the identification of functional sites. Note that the extraction is performed without a priori knowledge of the sequences. That implies the method has an ability for the automatic extraction of the signal patterns.

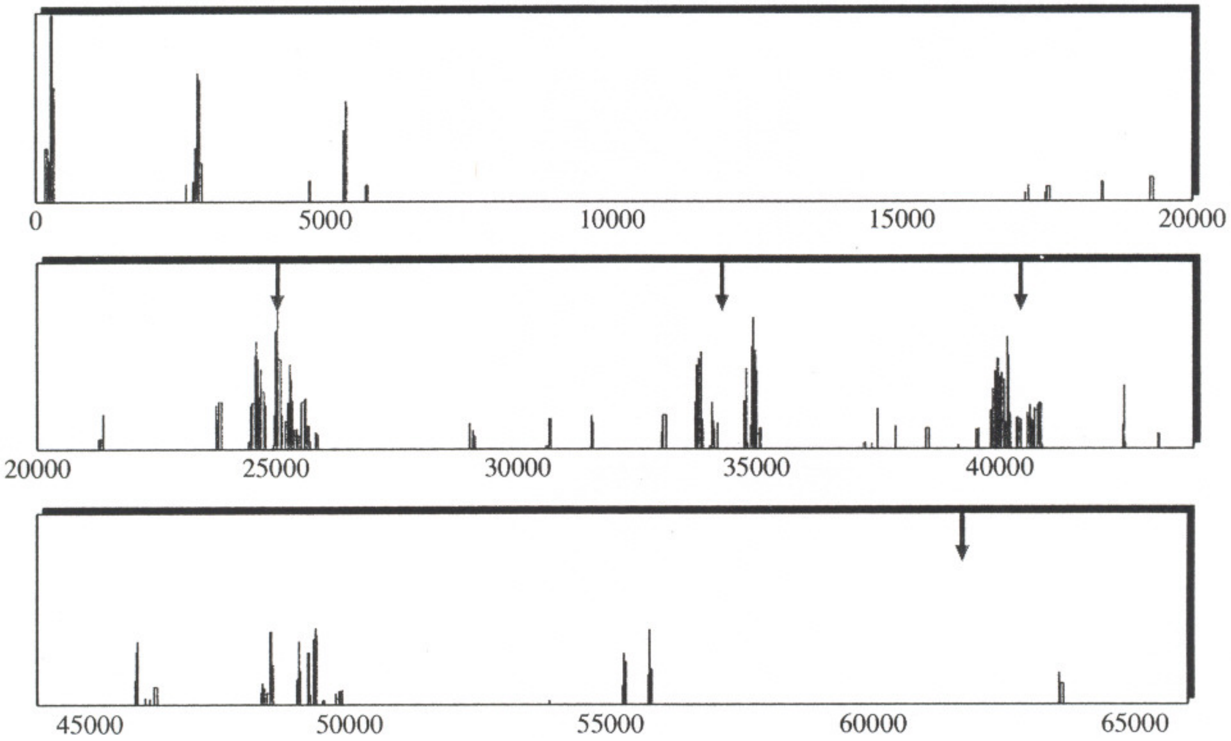


$\lambda = 1.00$ [$\log L(\hat{\theta}_i; f) = -26.518,\ p_i = 6$]

$\lambda = 0.25$ [$\log L(\hat{\theta}_i; f) = -25.995,\ p_i = 7$]

Figure 9: HMMs of 5' end of yeast introns

HSMHCAPG



HUMGHCSA

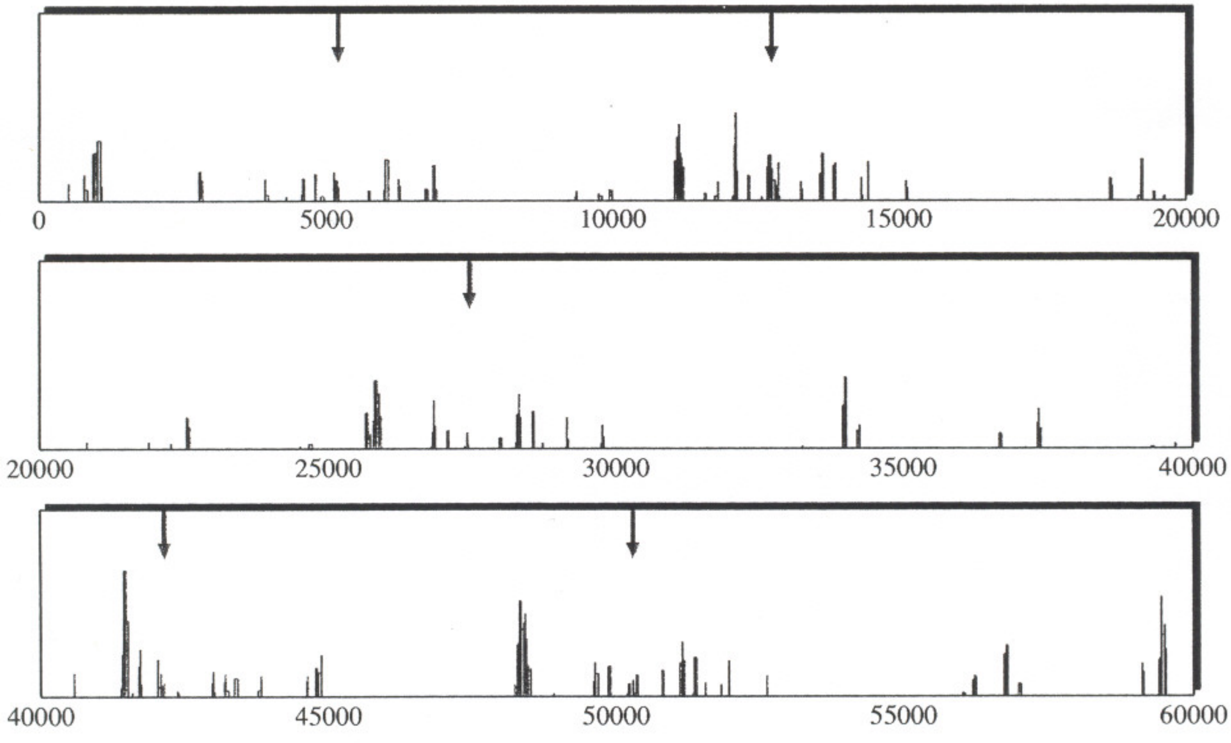


Figure 10: Recognition of promoter region

5 Conclusion

We have developed the method, for the extraction of the HMM representations of the signal patterns in DNA sequences. The GA has been used to design the optimal networks of the HMMs. The balance of likelihood and complexity of the models has been evaluated based on AIC. Using the method, the signal patterns of CAAT box, TATA box, promoter data set and 5' ends of introns have been extracted, and reasonable HMM representations have been obtained. The HMM representation of the promoter data set has shown a good potential for the recognition of the signal patterns. To conclude from the results of signal pattern extraction and recognition, the method is capable of extracting the patterns from given sequences and providing a reliable model for the identification of them, without a priori knowledge of the sequences.

Acknowledgments

The authors are grateful to Dr. Akihiko Konagaya, C&C Research Laboratories of NEC Corporation, for helpful comments on genetic operations in Genetic Algorithm.

References

[Akaike 73] Akaike, H: Information Theory and an Extension of the Maximum Likelihood Principle, In *Proc. of the 2nd Int. Symp. on Information Theory*, 267-281 1973.

[Asai *et al.* 93] Asai, K. *et al.*: Secondary Structure Prediction by Hidden Markov Model, *CABIOS*, **9**, 141-146 1993.

[Bucher 90] Bucher, P.: Weight Matrix Descriptions of Four Eukaryotic RNA Polymerase II Promoter Elements Derived from 502 Unrelated Promoter Sequences, *J.Mol.Biol.*, **212**, 563-578 1990.

[Fujiwara *et al.* 94] Fujiwara, Y. *et al.*: Stochastic Motif Extraction Using Hidden Markov Model, In *Proc. ISMB94*, 121-129 1994.

[GenBank 93] GenBank, Genetic Sequence Data Bank, Release 76.0, BBN Laboratories, U.S.A. (1993)

[Goldberg 89] Goldberg, D.: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[Haussler *et al.* 93] Haussler,D. *et al.*: Protein modeling using hidden Markov Models: Analysis of globins, In *Proc. HICSS26*, 792-802 1993.

[Holland 92] Holland, J.: *Adaptation in Natural and Artificial Systems*, MIT Press, 1992.

[Konagaya and Kondo 93] Konagaya, A. and Kondo,Y.: Stochastic Motif Extraction using Genetic Algorithm with the MDL Principle, In *Proc. HICSS26*, 746-755 1993.

[Konopka 93] Konopka, A. K.: Plausible Classification Codes and Local Compositional Complexity of Nucleotide Sequences, In *Proc. of the 2nd Int. Conf. on Bioinformatics, Supercomputing and Complex Genome Analysis*, 69-87 1993.

[Levin 94] Levin, B.: *Genes V*, Oxford University Press, 1994.

[Levinson *et al.* 83] Levinson, S. E. *et al.*: An Introduction to the Application of the Theory of Probabilistic Function of a Markov Process to Automatic Speech Recognition, *Bell Syst. Tech. J.*, **62**, 1035-1074 1983.

[Tanaka *et al.* 93] Tanaka, H. *et al.*: Classification of Proteins via Successive State Splitting of Hidden Markov Network, In *Proc. W26 in IJCAI93*, 1993.