Integrating Multiple Evidences by Hidden Markov Models

Kiyoshi Asai [1] **Katunobu Itou** [1] **Tetsushi Yada** [2]
`asai@etl.go.jp` `kito@etl.go.jp` `yada@mri.co.jp`

[1] Genome Informatics Group, Electrotechnical Laboratories, 1-1-4 Umezono, Tsukuba, 305-8568, Japan
[2] Mitsubishi Research Institute 2-3-6 Otemachi, Chiyoda-ku, 100-8141, Japan

1 Introduction

Hidden Markov Models (HMMs) have been used in genome informatics in many ways. In most cases, the output symbols of the HMMs are the four letters of nucleotide acids [2, 4, 9] or the twenty letters of amino acids [1, 7, 6]. However, we can build HMMs which have the other kinds of output symbols. We propose a new method which combine the sequence information and other pre-processed information by using multi-stream HMMs.

2 Output Symbols

An HMM is a stochastic signal source, whose transitions of hidden states belong to Markov process. The output symbols of this signal source can be any symbols, including discrete symbols, real values, real valued vectors, and multi streams of those values [1, 8, 5, 3]. The hidden states correspond to physical or conceptual labels, such as secondary structures of protein, coding regions of the genes, and the transcriptional signals in DNA. They are often the targets of the pattern recognition problems in genome informatics.

The output symbols of HMMs are the signals which we can observe. In the research of genome informatics, it is natural to set them to four nucleotide acids or twenty amino acids. However, the output symbols can be anything, as far as we can attach probability distributions of the output symbols.

In speech recognition, it is already known the almost-best pre-process for the speech signal (wave form), Fourier type pre-process and their differentials.

In the case of DNA/protein sequences, we have not yet found the best pre-process of the sequences, but we know some useful pre-process. We can pre-process the DNA/protein sequences and set the output symbols of HMMs to the results of the pre-process. For example, we can use, as symbols, triplets or any length of the sequences, Fourier transform of the sequences, hydrohobicity of the regions. It is known that those values are important in gene identification from DNA sequences and in secondary structure prediction of protein.

3 Multiple Stream

A multi-stream HMM is an HMM which outputs multiple streams of output symbols. We can assign a preprocessor for each stream, where we can regard these preprocessors the 'views' of the sequences. The preprocessors can be Fourier transform, grouping of k-tuples, scores of homology search, coding potentials, hydrohobicity etc. The parsing of the multi-stream HMM is straight forward because the logarithm of joint output probabilities are calculated as the sum of the logarithm of output probabilities of the streams. By doing so, we can combine different types of evidences hidden in the DNA/protein sequences into the stochastic parsing and find the most probable understanding of the sequences. We will show the result of using the multi-stream HMMs for the gene recognition system.

Acknowledgments

This work was supported in part by a Grant-in-Aid (08283101:"Genome Science") for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture of Japan, and in part by "Genome Informatics" project of Agency of Industrial Science and Technology, The Ministry of International Trade and Industry of Japan. The authors thank Dr. Otsu, the director of Machine Understanding Division and the members of Genome Informatics Group of Electrotechnical Laboratories for the support and the discussions.

References

[1] Asai, K., Handa, K. and Hayamizu, S., Genetic Information Processing by Stochastic Model: HMM for Secondary Structure Prediction of Protein, *Genome Informatics,* 2:144-147 (in Japanese), 1991.

[2] Asai, K., Yada, T. and Itou, K., Finding Genes by Hidden Markov Models with a Protein Motif Dictionary, *Genome Informatics 1996,* Universal Academy Press, 88-97, 1996.

[3] Asai, K., Yada, T. and Itou, K., Automatic Gene Recognition without Using Training Data, *Genome Informatics 1997,*, Universal Academy Press, 15–24, 1997.

[4] Asai, K., Ueno, Y, Itou, K. and Yada, T., Recognition of Human Genes by Stochastic Parsing, *PSB98*, World Scientific, 228–239, 1998.

[5] Burge, C. and Karlin, S., Prediction of Complete Gene Structures in Human Genomic DNA, *J. Mol. Biol.*, 268:78–94, 1997.

[6] Fujiwara, Y., Asogawa, M. and Konagaya, A., Stochastic Motif Extraction Using Hidden Markov Model, *ISMB94*, AAAI Press, 121–129, 1994.

[7] Haussler, D., Krogh, A., Mian, I.S. and Sjölander, K., Protein modeling using hidden Markov Models: Analysis of globins, *26th HICSS*,792–802. (1993).

[8] Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H., Integrating Database Homology in a Probabilistic Gene Structure Model, *PSB97*, World Scientific, 232–244, 1996.

[9] Yada,T. *et al.*, Extraction of Hidden Markov Model Representations of Signal Patterns in DNA Sequences, *PSB96*, 686–696, 1996.