Gene Finding Using HAKKE System

Kiyoshi Noda [1] **Satoshi Matsumoto** [1] **Ayumi Shinohara** [1]
Takayoshi Shoudai [1] **Satoru Miyano** [2]

{knoda, matumoto, ayumi, shoudai}@i.kyushu-u.ac.jp
miyano@ims.u-tokyo.ac.jp

[1] Department of Informatics, Kyushu University
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-81, Japan

[2] Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

1 Introduction

We are developing a machine learning system HAKKE, that is a hybrid system cooperating with algorithms of a pool. HAKKE produces a prediction algorithm, which is suitable for predicting functional regions from sequences. It employs an extension of the weighted majority algorithm for adjusting the weights of algorithms. One of the advantages of such a multi-strategy system is the possibility of finding a better predictor than any predicting algorithm in the pool. Actually, by the experiments on transmembrane domain sequences and α-helix predictions, we have verified that the accuracy of the predictor produced by HAKKE is much higher than that of any prediction algorithms in the pool [4].

In Genome Informatics, one of the most important and challenging problem is to detect genes in DNA sequences. A number of gene finding systems have been proposed so far, such as GENMARK [1], GRAIL [3], GeneHacker [5], GENSCAN [2]. In this paper, we try to combine these systems using HAKKE, in order to get a more accurate gene finding system. We report some preliminary results on the experiments.

2 HAKKE system

HAKKE employs an extension of the weighted majority algorithm (WM) as a core of the system. WM assumes a pool of prediction algorithms, each of which answers 0 or 1 for any question. Initially, WM assigns a positive weight to each algorithm of the pool. WM makes its prediction by weighted majority voting of the algorithms. When the prediction of WM was incorrect, WM gives a penalty to each algorithm which voted incorrectly: the penalty is done by decreasing the weight by multiplying a fixed real number β $(0 \leq \beta < 1)$. Our extension of WM is to introduce an "abstention" in the voting. Namely, each algorithm in the pool is permitted to answer "I do not know" to the questions for which it has little confidence (Fig. 1).

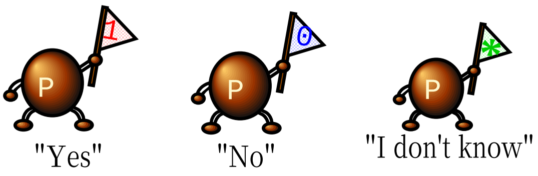

Figure 1: Abstention

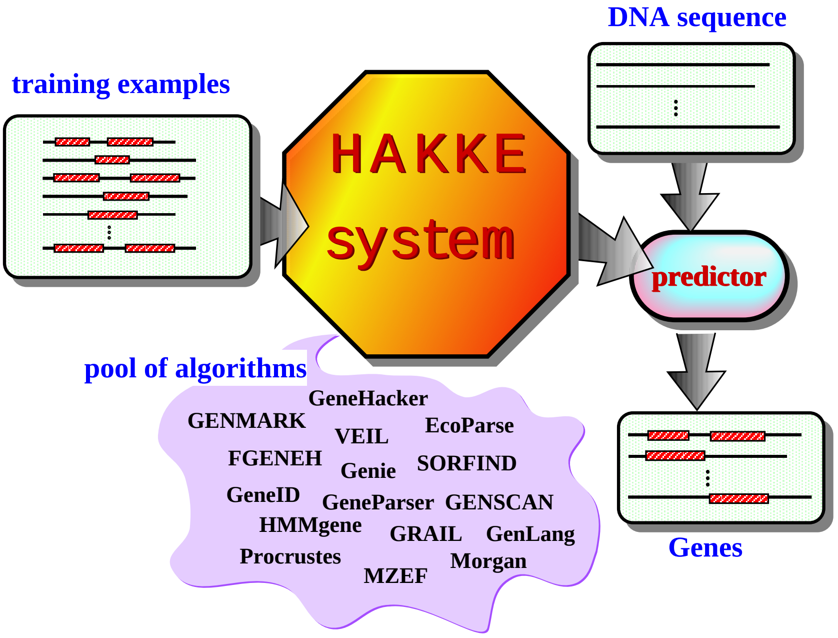

Figure 2: Gene Finding using HAKKE system

As input, HAKKE takes a set of marked examples for predicting a marking ϕ of an unknown sequences. After a number of weighted majority votes, the system produces a predictor which approximates the marking ϕ of the unknown sequences.

3 Method of Experiments

We select several well-known gene finding system, such as GENSCAN [2], GRAIL [3], predicting algorithms of a pool (Fig. 2). HAKKE accesses to these systems through Internet, by using Perl scripts. We apply the system to human genes. We are now in the working process, and we will report the experimental results at the conference site.

Acknowledgments

The authors thank to Katsuhiko Murakami for his great help to write perl scripts.

References

[1] M. Borodovsky and J. McIninch, "GENMARK: parallel gene recognition for both DNA strands," *Comp. Chem.* Vol. 17, pp. 123–133, 1993.

[2] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *J. Mol. Biol.*, Vol. 268, pp. 78–94, 1997.

[3] E.C. Uberbacher and R.J. Mural, "Locating Protein-Coding Regions in Human DNA Sequences by a Multiple Sensor-Neural Network Approach," *Proc. Natl. Acad. Sci USA*, Vol. 88, pp. 11261-11265, 1991.

[4] N. Furukawa, S. Matsumoto, A. Shinohara, T. Shoudai and S. Miyano, "HAKKE: A Multi-Strategy Prediction System for Sequences," *Genome Informatics 1996*, 98–107, 1996.

[5] T. Yada and M. Hirosawa, "Gene Recognition in Cyanobacterium Genomic Sequence Data Using the Hidden Markov Model," *ISMB'96*, Vol. 4, pp. 252–260, 1996.