Automatic Discovery of Hidden Markov Representations for Functional Sites within DNA Sequences

Tetsushi Yada [1]
yada@tokyo.jst-c.go.jp

Yasushi Totoki [2]
totoki@idas.imslab.co.jp

Masato Ishikawa [3]
ishikawa@trl.mei.co.jp

Kiyoshi Asai [4]
asai@etl.go.jp

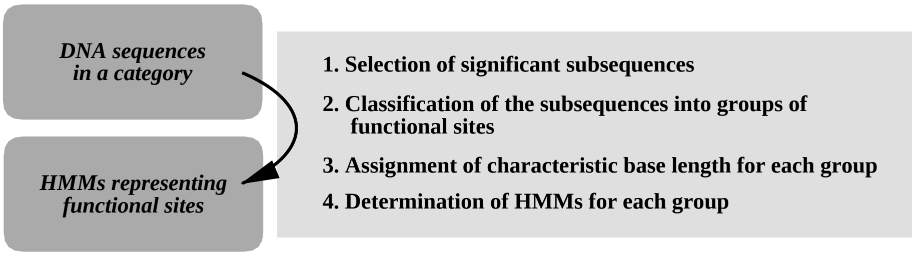
[1] Japan Science and Technology Corporation (JST)
5-3 Yonbancho, Chiyoda-ku, Tokyo 102, Japan

[2] Information and Mathematical Science Laboratory, Inc.
2-43-1 Ikebukuro, Toshima-ku, Tokyo 171, Japan

[3] Matsushita Electric Industrial Co.,Ltd.
4-5-15 Higashi-shinagawa, Shinagawa-ku, Tokyo 140, Japan

[4] Electrotechnical Laboratory (ETL)
1-1-4 Umezono, Tsukuba 305, Japan

We have developed a fast and sensitive method to automatically discover representation models for functional sites within DNA sequences. In the method, functional sites are represented as a stochastic model called hidden Markov model (HMM) [1] whose network topology is left-to-right. HMM has some advantages of description over conventional representation models such as weight matrix [2] and regular expression [3]. The method consists of the following steps (Fig. 1): (1) selection of significant subsequences; (2) classification of the subsequences into groups of functional sites; (3) assignment of characteristic base length for each group; (4) determination of HMM for each group. We have applied the method to automatic discovery of hidden Markov representations for functional site within human promoter sequences. The method discovered 40 representation models with characteristic base lengths. Some of them corresponded to TATA box (Fig. 2), CAAT box (Fig. 3), GC box, Cap site, CT signal, Octamer, κB and ATF. To validate the method, we have applied the models to recognition of functional sites within the promoter sequences. As results, recognition accuracy of TATA, CAAT and GC box were 85.4, 75.8 and 77.2%, respectively. This indicates that good representation models for functional sites has been obtained by the method.


DNA sequences in a category
HMMs representing functional sites

1. Selection of significant subsequences
2. Classification of the subsequences into groups of functional sites
3. Assignment of characteristic base length for each group
4. Determination of HMMs for each group

Figure 1: Schematic diagram of the method

Acknowledgement

This work was carried out as a part of ALIS (Advanced Lifescience Information System) project for genome analysis by Japan Science and Technology Corporation (JST).

References

[1] A. Krogh, M. Brown,. I. S. Mian, K. Sjölander and D. Haussler, "Hidden Markov Models in Computational Biology, Applications to Protein Modeling", *J. Mol. Biol.*, Vol. 235, pp. 1501-1531, 1994.

[2] M. Gribskov, R. Lüthy and D. Eisenberg, D, "Profile Analysis", In *Methods Enzymol*, Vol. 183, pp. 146-159, 1990.

[3] A. Bairoch, "PROSITE: A Dictionary of Sites and Patterns in Protein", *Nucleic Acids Res.*, Vol. 20, pp. 2013-2018, 1992.

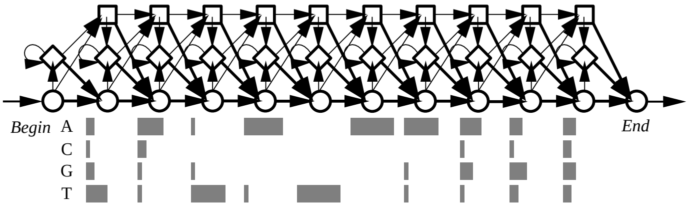


Figure 2: An HMM representing TATA box

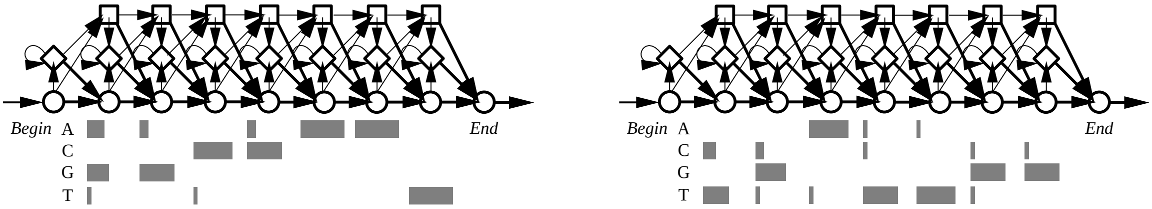


Figure 3: HMMs representing CAAT box