GeneHacker: Gene–finding Program for the Prediction of Precise Protein Coding Regions

Makoto Hirosawa [1]
hirosawa@kazusa.or.jp

Tetsushi Yada [2]
yada@tokyo.jst-c.go.jp

[1] Kazusa DNA Research Institute,
1532–3 Yana-uchino, Kisarazu, Chiba 292, JAPAN

[2] Japan Science and Technology Corporation(JST),
5–3 Yonbancho, Chiyoda–ku, Tokyo 102, JAPAN

1 Introduction

Recent world-wide efforts in large-scale genomic sequencing have accelerated the production of long contiguous nucleotide sequence data. Consequently, reliable methods, which require less human judgment, of gene-identification in such nucleotide sequence data are desired. For the analysis of prokaryotic sequences, GeneMark[1] based on Markov model, has been widely used. The coding regions in the whole nucleotide sequence of *Haemophilus influenzae* Rd., *Mycoplasma genitalium*, Cyanobacterium *Synechocystis* sp. strain PCC6803[2] and *Methanococcus jannaschii* were assigned by GeneMark together with similarity search. GeneMark utilizes two kinds of statistics, those of coding regions and those of noncoding regions, represented in the framework of the Markov model to predict coding regions with plural candidates of their translation initiation sites. The selection of one translation initiation site for a coding region among candidates is up to users. Normally, the translation initiation sites which make coding regions the longest are selected.

However, for the understanding of the biochemical process in living organisms, the assignment of protein coding regions with precise translation initiation sites is important. For the assignment of precise translation initiation sites, two kinds of things are required. One is a paradigm which represents information necessary for the precise assignment. The other is the extraction of the information.

Representation of the necessary information other than the statistics of coding regions and noncoding regions, such the Shine-Dalgarno sequence, is possible by using HMM (the Hidden Markov Model), which is the higher paradigm of the Markov model and has more flexible

and extensible than the Markov model. Investigation of gene-finding under the paradigm of HMM have been studied by Krogh *et al.* (*Escherichia coli*)[3] and by Yada and Hirosawa (Cyanobacterium *Synechocystis* sp. strain PCC6803)[4]. The prediction rates of protein coding regions at least comparable to those by GeneMark have been achieved by the both study.

For the extraction of the information around translation initiation sites, enough number of experimentally determined translation initiation sites of protein coding regions is indispensable. Recently, Sazuka and Ohara determined translation initiation sites of 72 proteins by two dimensional electrophoresis and assigned their protein coding regions exactly in the genome of the above mentioned cyanobacterium[5]. Hirosawa *et al.* have succeeded in the designing of an HMM representing sequences around translation initiation sites based on the 72 sequences (in Preparation).

2 Precise prediction of protein coding regions

Taking the cyanobacterium as a model organism[2] we have developed the HMM which contains information necessary for the precise prediction of protein coding regions in addition to the information on coding regions and noncoding regions. The basic structure of the HMM was essentially the same as the HMM previously developed by Yada and Hirosawa[4]. The HMM representing the Shine-Dalgarno sequence, extracted by Hirosawa *et al.*(in Preparation) is attached to the basic HMM. We are now engaged in the analysis of the protein coding regions predicted with the described HMM. Application of our method to other organisms, such as *E.coli* and *B.subtilis* are planed after completion of our analysis on the cyanobacterium.

References

[1] Borodovsky M. and McIninch J.D. (1993). GENMARK:Parallel gene recognition for both DNA strands. *Computer Chemistry* 17, 123-133.

[2] Kaneko, T., Sato, S., Kotani, H., et al. (1996). Sequence analysis of the genome of the unicellular Cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions.*DNA Res.*, 3, 109-136

[3] Krogh,A., Brown.M., Mian,I.S., Sjölander,K. and Haussler,D. (1994) Hidden Markov Models in Computational Biology, Applications to Protein Modeling, *J. Mol. Biol.* **235:**1501–1531.

[4] Yada, T., and Hirosawa, M. (1995). Recognition in cyanobacterium genomic sequence data using the Hidden Markov Model. *Proceedings of International Conference on Intelligent System for Molecular Biology (ISMB-96)*, 252-260.

[5] Sazuka,T. and Ohara,O. 1996, Sequence features surrounding the translation initiation sites assigned on the genome sequence of *Synechocystis* sp. strain PCC 6803 by amino-terminal protein sequencing, *DNA Res.*, **3** (In Press).