Prediction Rate of Coding Regions is Enhanced upto 99.15 % by Joint Use of GeneMark-RC and GeneHacker in Case of a Cyanobacterium

Makoto Hirosawa
hirosawa@kazusa.or.jp
Kazusa DNA Research Institute
1532-3 Yana, Kisarazu-shi, Chiba 292, Japan

1 Studies on coding region assignment

The advancement in large-scale sequencing has accelerated the production of long contiguous nucleotide sequence data. The whole genomic sequence data is currently available for several prokaryotic organisms. The first step in the analysis of genomic sequence data is to assign coding regions, which is absolutely necessary for a comparative study of one organism with the others and to elucidate common as well as specific features among them. For coding region assignment, two kinds of problems must be solved. One is the detection of coding regions, and the other is precise assignment of translation initiation sites of coding regions. I have studied to solve them with collaborators by taking the cyanobacterium *Synechocystis* sp. strain PCC6803 [1] as a model organism. For the study of the precise assignment, HMM (hidden Markov model) was selected as theoretical framework, and the analysis and assignment of initiation sites were studied for highly expressed genes and for photosynthetic genes. For the study of coding region detection, two kinds of programs (or procedure), GeneHacker [2] and GeneMark-RC [3] have been developed. In this presentation, effectiveness of joint use of the two programs is demonstrated.

2 GeneMark-RC and GeneHacker have their merits and drawbacks

GeneMark, developed by Borodovsky *et al.*, has been widely used for coding region assignment [4]. The GeneMark program identifies coding regions based on the statistical properties of nucleotide permutation within coding regions that differ from those observed in non-coding regions. The necessary statistics for GeneMark are described in term of the Markov model. With recent advance in computer performance it comes to feasible to apply HMM, a more advanced concept of the Markov model for gene–finding. Yada and I have developed GeneHacker, a gene–finding program based on HMM. Specific feature of GeneHacker is its use of di–codon statistics. Prediction rate of GeneHacker (92.9 %) was proved to be slightly better than that of GeneMark (91.9 %) in case of 1.0 Mb region of the species. GeneHacker can detect short coding regions better than GeneMark mainly due to its use of *hidden* Markov model at the expense of computational time.

Borodovsky *et al.* showed that an increase in GeneMark accuracy can be achieved by deriving specific Markov models for groups (classes) of genes that differ in their characteristics in nucleotide permutation [5]. In the case of *E. coli*, genes can be divided into three classes using the clustering method termed *correspondence analysis* [6]. Based on the classification, they derived three class-specific GeneMark matrices. The prediction rate of coding regions was distinctly improved by the introduction of these class-specific matrices [5].

Recently, in collaboration with Borodovsky *et al.*, I have developed GeneMark-RC, a GeneMark-based recursive procedure for the identification and classification of coding regions in genomic sequence

data [3]. Prediction rate of 98.3 % was accomplished for the 1.0 Mb region. Unlike the case in *E. coli*, the classification was done with its application to GeneMark in consideration.

3 Drawbacks can be remedied by their joint use

Drawbacks of GeneMark-RC (or GeneMark) and GeneHacker can remedied by using the both methods jointly. Consequently, the prediction rate can be enhanced. By taking the whole genomic region of the species as an example [7, 8], complementary nature of the two methods was demonstrated.

The whole process of classification and detection of ORFs by GeneMark-RC was completed within three hours with Ultra 2 (Creator Model 1300) of Sun Microsystems. With three kinds of statistics corresponding to three derived class and one additional statistics representing authentic genes of the species, 3103 of 3168 annotated ORFs were detected [7]. Therefore, its prediction rate was 97.9 %. Detection of ORFs by GeneHacker was completed within four days with Dec Alpha station. Prediction rate by GeneHacer was 96.8 % (3068/3168) %. GeneHacker couldn't detect some ORFs which were detected by GeneMark-RC with Class 2 statistics largely contributed from exogenous genes. However, GeneHacker served to detect short ORFs which couldn't be detected by GeneMark-RC.

ORFs escaped detection of the both methods was 27, therefore the false negative error rate by the joint method was as low as 0.85 %. Among the annotated ORFs, there are ORFs which were solely assigned based on their length, and some of them might be spuriously assigned. Therefore, actual false negative error may be much lower.

References

[1] Kaneko, T., Sato, S., Kotani, H. *et al.* 1996, Sequence Analysis of the Genome of the Unicellular Cyanobacterium *Synechocystis* sp. strain PCC 6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, *DNA Res.,* **3,** 109–136.

[2] Yada, T. and Hirosawa, M. 1996, Recognition in cyanobacterium genomic sequence data using the Hidden Markov Model, *in the Proceedings of International Conference on Intelligent System for Molecular Biology (ISMB-96),* 252–260.

[3] Hirosawa, M., Isono, K., Hayes, W.S. and Borodovsky, M. Gene Identification and Classification in the Synechocystis Genomic Sequence by Recursive GeneMark Analysis, *DNA Sequence*, *In Press.*

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

[5] Borodovsky, M., McIninch, J.D., Koonin, E.V, Rudd, K.E., Medigue, C. and Danchin, A. 1995, Detection of new genes in a bacterial genome using Markov Models for three gene classes. *Nucleic Acids Research*, **23**, 3554–3562.

[6] Medigue, C., Viari, A., Henaut, A. and Danchin, A. 1993, Colibri: a functional database for the *Escherichia coli* genome. *Microbiological Review*, **57**, 623–654.

[7] Hirosawa, M. and Isono, K. 1997, GeneMark-RC, a recursive procedure with self-consistency evaluation for the detection and classification of ORFs; its application to the analysis of prokaryotic genomes, *in the Proceedings of the Eighth Workshop on Genome Informatics*, In Press.

[8] Yada, T. 1997, Gene Identification in Prokaryotic Genomes using Hidden Markov Model. *PNE*, in Press.