Modeling Transcriptional Units of *E. coli* Genes Using HMM

Tetsushi Yada [1] yada@mri.co.jp
Mitsuteru Nakao [2] nakao@kuicr.kyoto-u.ac.jp
Yasushi Totoki [3] totoki@imslab.co.jp
Takahiro Ishii [4] ishii@imcb.osaka-u.ac.jp
Kenta Nakai [4] nakai@imcb.osaka-u.ac.jp

[1] Mitsubishi Res. Inst., Inc., 2-3-6 Otemachi, Chiyoda-ku, Tokyo 100-8141, Japan
[2] Institute for Chemical Research, Kyoto Univ., Gokasho, Uji, Kyoto 611-0011, Japan
[3] Information and Mathematical Science Lab., Inc., Ikebukuro Aoyagi Bldg., 2-43-1, Ikebukuro, Toshima-ku, Tokyo 171-0014, Japan
[4] Inst. Mol. Cell. Biol., Osaka Univ., 1-3 Yamada-oka, Suita, Osaka 565-0871, Japan

1 Introduction

In recent years, the number of bacteria whose entire genomic sequence is determined is growing rapidly. However, the information which can be derived from computer analyses of them is still limited although the predictive identification of coding regions is performed with relatively high accuracy (see [4], for example). Therefore, we have studied ways to interpret the regulatory information coded in genomic sequences [5, 6]. In this work, we report our first effort to integrate the models for detecting various signals (*e.g.*, promoters and terminators) with our previous model of coding regions, aiming at the recognition of transcriptional units in bacterial genomes.

2 Model

Roughly speaking, our model consists of the parts which recognize encoded signals (*i.e.*, promoters, terminators, and ribosome-binding sites), the part which recognizes coding regions, and the parts which models the spacer regions between these elements. The entire model is constructed by the scheme of hidden Markov model (HMM). One minor exception is the treatment of terminators because the stem-loop structure of rho-independent terminators is difficult to directly represent in HMM. To avoid this difficulty, we calculated the probability that rho-independent terminators exist at each position before the parsing. We used the genome sequence of *E. coli* by Blattner *et al.* [1]. For the model of promoters, only the ones which are dependent on sigma 70 are considered in this stage. The model was derived from the multiple alignment of 441 promoters compiled by Ozoline *et al.* [3]. For terminators, we have only implemented the model which are independent from the rho-factor. It was built using the 145 positive data compiled by Carafa *et al.* [2], and arbitrarily compiled 4176 negative data. To construct the model of ribosome-binding site, the upstream 25 bp segment from the start codon was taken from each of the 4287 coding regions and they were subsequently aligned. The model of coding regions was built based on the di-codon statistics of all coding regions and the control data of non-coding regions. The model of spacer regions is constructed using the length distribution and base contents between two neighboring elements. Finally, our HMM tries to find the most consistent model in terms of spacers, logical orders and likelihood of three signals, and the probability of finding coding regions.

3 Results and Discussion

To assess the plausibility of our model, a collection of *E. coli* annotated operon structures by Blattner *et al.* [1] was used. As the control of our model, a simple prediction scheme was introduced; that is, given the complete set of coding sequences, a set of consecutive coding sequences all of which are on the same strand and all neighboring pairs do not have an inter-coding region exceeding the cutoff length. By varying the cutoff value, we obtained its optimum value, about 30 bp, and the corresponding recognition accuracy was 53.3%. On the contrary, our model showed a rather poor value, 43.3%. Of course, it is a very preliminary result and we are optimistic in beating the control result. In this stage, it should be noted that the prediction of operon structure is not trivial at all. We plan to add the model of rho-dependent terminators, soon. Later, we should also add the knowledge of various transcription factors including sigma factors and compare the results with experimental data produced from post-sequencing projects.

Acknowledgements

This work was supported in part by Special Coordination Funds for Promoting Science and Technology from Science and Technology Agency. KN was also supported by a Grant-in-Aid for Scientific Research on Priority Areas, Genome Science, from the Ministry of Education, Science, and Culture of Japan and by Research for the Future Program of JSPS.

References

[1] Blattner, F.R., Plunkett, G. 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shao, Y., The complete genome sequence of *Escherichia coli* K-12, *Science*, 277:1453–1474, 1997.

[2] Carafa, Y.d'A., Brody, E., and Thermes, C., Prediction of Rho-independent *Escherichia coli* transcription terminators: a statistical analysis of their RNA stem-loop structures, *J. Mol. Biol.*, 216:835–858, 1990.

[3] Ozoline, O.N., Deev, A.A., and Arkhipova, M.V., Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase, *Nucl. Acids Res.*, 25:4703–4709, 1997.

[4] Yada, T. and Hirosawa, M., Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model, *DNA Res.*, 3:355–361, 1996.

[5] Yada, T., Totoki, Y., Ishii, T., and Nakai, K., Functional prediction of *B. subtilis* genes from their regulatory sequences, *Proc. Intell. Syst. Mol. Biol.*, 5:354–357, 1997.

[6] Yada, T., Totoki, Y., Ishikawa, M., Asai, K., and Nakai, K., Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences, *Bioinformatics*, 14:317–325, 1998.