GeneHacker Plus: Yet Another Gene-Finding Program with Top-Class Accuracy in Locating Start Codons

Tetsushi Yada[1]
yada@gsc.riken.go.jp
Yasushi Totoki[1]
totoki@gsc.riken.go.jp
Toshihisa Takagi[2]
takagi@ims.u-tokyo.ac.jp
Kenta Nakai[2]
knakai@ims.u-tokyo.ac.jp

[1] Genomic Sciences Center, RIKEN, c/o Laboratory of Genome Database, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan
[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Keywords: gene finding, bacterial genome analysis, translational start site

1 Introduction

The entire genomes of various bacterial species continue to be sequenced even in a faster pitch. Therefore, there is a great need for bacterial gene-finding programs, *i.e.*, programs for locating plausible coding regions from given genomic sequences. So far, many programs have been developed. One of the authors' group also developed such a program, GeneHacker [4]. However, it has been pointed out that these programs are practically useful in finding the presence of coding regions but are still inaccurate in locating the position of start codons precisely. One reason is that many of the annotations of coding regions have not been experimentally verified. Recently, a powerful bacterial gene-finding program, GeneMark.hmm, was released [2]. Not only it is probably the most reliable program but also it shows higher accuracy in locating start sites when assessed with smaller data obtained from a proteome project. Hannenhalli *et al.* also developed a specialized method for predicting bacterial translational start-sites and payed great efforts in assessing its accuracy by using rather reliable data [1]. But their method was not designed as a module of gene-finding programs and combines various features of potential start sites. Here, we report our novel program, GeneHacker Plus. In spite of the simplicity of its algorithm, its performance is remarkable as we briefly describe below.

2 Methods

Like many other gene-finding programs, GeneHacker Plus uses the technique of hidden Markov model (HMM) to model the gene structure. More specifically, it is an 'HMM with duration' that can deal with the length distribution of any regions.

Our treatment of coding potentials, *i.e.* statistical features used for the discrimination of true coding regions from ORFs, is basically similar to that of GeneMark.hmm; it divides genes into two classes, namely 'typical' genes and 'atypical' genes and use the di-codon statistics for each of them. However, the algorithm for this classification is different; we formulate an iterative algorithm that can be applied to any bacterial species. It turned out that about 10% to 20% of genes are classified into the 'atypical' class.

The treatment of the sequence features around the translational start sites is the key in our method. One of the difficulties is that such features can vary with different species. For example, in

Synechocystis species, only a weak consensus sequence for the ribosome-binding site can be found and the nucleotides of some positions around the start site are conserved [3]. To consider such variations systematically, we constructed a model of the upstream region as follows: first, we took the 25bp regions from a set of start sites in training data and aligned them prohibiting inner gaps; second, the most significantly conserved subregion (expected to be the ribosomal binding site) was detected by the χ^2 test; third, by aligning the remaining regions at their 3' end, significantly conserved positions are sought expecting to detect the positions near the start codon in *Synechocystis*; fourth, the remained positions are modeled as variable regions. The model of the conserved regions were constructed based on the conditional probability of two neighboring residues unless the region consists of one position. The model of the remaining variable region was an HMM based on the dinucleotide statistics and the length distribution.

3 Performance of GeneHacker Plus

We assessed the performance of GeneHacker Plus by several ways. First, we compared the performance in recognizing the coding regions used in the training data since corresponding results are reported for GeneMark.hmm [2]. Although both programs could predict the position of true stop codons rather accurately, GeneHacker Plus seems to outperform GeneMark.hmm in its specificity. For example, for *E. coli* genes, GeneHacker Plus showed 96.4% sensitivity and 98.1% specificity while GeneMark.hmm showed 97.3% sensitivity but only 91.8% specificity. It seems that these values do not vary drastically for predicting new genes; in the ten-fold cross-validation test, GeneHacker Plus showed 95.9% sensitivity and 98.1% specificity. In addition, we confirmed that four out of seven recently-identified *B. subtilis* genes had been treated as false positives because they have not been included in the training data.

Second, the accuracy in exactly pinpointing the start codons was compared with that of Hannenhalli *et al.*'s method [1] (Hannenhalli, personal comm.). Although GeneHacker Plus constructed the model of upstream region in a rather simple way and does not assume the presence of coding regions for testing, its performance seems to be comparable with or better than that of the specialized method; for example, in the cross-validation test, our method could detect 95.7% of experimentally-detected 184 start sites of *E. coli* while theirs could detect 84.9% of them.

Lastly, we also pursued the possibility of raising the prediction accuracy by incorporating the homology information. The results of BLAST searches were directly embedded in our hidden Markov model. Generally speaking, a discovery of significant similarity can have a greater value because it is not a result of mere 'prediction'. Therefore, we do not deny the potential value of this option. However, such homology searches are rather time-consuming. Moreover, we could not observe significant improvements in the prediction accuracy.

References

[1] Hannenhalli, S.S., Hayes, W.S., Hatzigeorgiou, A.G., and Fickett, J.W., Bacterial start site prediction, *Nucl. Acids Res.*, 27(17):3577–3582, 1999.

[2] Lukashin, A.V. and Borodovsky, M., GeneMark.hmm: new solutions for gene finding, *Nucl. Acids Res.*, 26(4):1107–1115, 1998.

[3] Sazuka, T., Yamaguchi, M., and Ohara, O., Cyano2Dbase updated: linkage of 234 protein spots to corresponding genes through N-terminal microsequencing, *Electrophoresis*, 20(11):2160–2171, 1999.

[4] Yada, T. and Hirosawa, M., Gene recognition in cyanobacterium genomic sequence data using the hidden Markov model, *Proc. 4th Int. Conf. Intell. Systems Mol. Biol.*, 252–260, 1996.