Finding Coding Region Using Secondary Hexamer Measure and Two-Dimensional Linear Discriminant Analysis

Katsuhiko Murakami [1] [2] Toshihisa Takagi [1]

[1] Human Genome Center
Institute of Medical Science, University of Tokyo
{katsu, takagi}@ims.u-tokyo.ac.jp

[2] Central Research Laboratory, Hitachi Ltd.

Abstract

We have developed a coding region prediction system. It is constructed from several measures that indicate exonness of a region in DNA sequence. The system includes a new statistical measure called secondary hexamer measure which we have developed. In addition to the measure, several measures are combined by two-dimensional linear discriminant analysis (2D-LDA). Then the system outputs a best gene model, that is a model with the best score accumulated by phase-specific dynamic programming. Our test of this program on 568 vertebrate complete gene sequences had 61% accuracy at exon level for exact match and 95% accuracy at nucleotide level. The average correlation coefficient (CC) between prediction and actual structure was 0.80.

1 Introduction

Human Genome Project facilitates production of huge amount of DNA sequence data. One of the main aims of the Human Genome Project is to make a catalog of genes, with their locations, products, functions, etc. The primary computational analysis for the newly determined sequences are database search technique such as BLAST and coding region prediction. Since not all DNA sequences are stored in databases, the prediction of protein coding regions based on statistical information is important, and have been developed in this decade. Though there are several coding region prediction systems, their correlation coefficients (CC) between prediction with the actual gene structure range from 0.65 to 0.88 for vertebrate DNA sequences[1]. This result indicates that these systems are still far from complete level, in which case CC equals to 1. To develop a more accurate gene identification system, many new techniques are required. Here, we describe a coding region prediction system constructed with a new measure. We also show the performance of the system.

2 Data and Methods

We have retrieved complete and partial human gene sequences from GenBank (Release 83, June 1994). It is required that the label of 'Homo sapiens' on the item 'ORGANISM'. Some data including contradiction are deleted. Furthermore, pseudo genes, genes with alternative splicing, entries with fusion nucleotides, are discarded. Of the remaining 353 data, we took two-thirds(235 entries, including 82 complete genes) as training data set.

Given a DNA sequence, the system first suppose possible exon candidates (first, internal, last exons) and calculates several exon measures, and make a score combining the measures by 2D-LDA. The measures are calculated by weight matrices, neural networks, in-frame hexamer, secondary hexamer, local complexity, and fourier coefficient. The secondary hexamer measure is defined the same way as traditional primary hexamer measure, except for its learning data. Special exons whose score are remarkably high by the traditional hexamer measure are deleted beforehand from the learning data. Therefore, the new measure learns the exons that are rarely appear in coding region. Finally the system construct a gene model with the best accumulated score using phase-specific dynamic programming.

3 Results and Discussion

Hexamer measure is the most widely used in coding region prediction systems. However it has a drawback that low GC content coding regions are difficult to detect[1, 2, 3, 4]. The secondary hexamer measure mainly reflects the statistical patterns of low GC content. It is studied that the discrimination power of fourier measure is less dependent of GC content[2]. To detect low GC content coding region, we have introduced secondary hexamer measure and used known fourier measure.

The performance of the system on 568 vertebrate complete gene sequences had 61% accuracy at exon level for exact match and 95% accuracy at nucleotide level. The average correlation coefficient (CC) between prediction and actual structure was 0.80. This result of CC is better than GRAIL(v1.3b, CC=0.79), GeneFinder or FGENEH (CC= 77) on this test. The other systems using homology search result, however, perform better than our system. The CC for Genie is presumed to be approximately 0.91, and 0.88 for GeneID+, and 0.85 for GeneParser3. This indicate the total analysis including both statistical approach and case based approach (homology search) is so effective that we should adopt homology search result and then evaluate the total performance in the future.

Acknowledgments

This work is partially supported by Grant-in-Aid for Scientific Research on Priority Areas, "Genome informatics" from the Ministry of Education, Science, Sports and Culture, Japan and by Science and Technology Agency, Japan.

References

[1] M. Burset and R. Guigo. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, 1996.

[2] R. Guigo and J.W. Fickett. Distinctive sequence features in protein coding genic non-coding and intergenic human DNA. *JMB*, 253:51–60, 1995.

[3] R. Lopez, F. Larsen, and H. Prydz. Evaluation of the exon predictions of the grail software. *Genomics*, 24:133–136, 1994.

[4] E.E. Snyder and G.D. Stormo. Identification of protein coding regions in genomic DNA. *JMB*, 248:1–18, 1995.