Reinforcement Learning for Improving Gene Identification Accuracy by Combination of Gene-Finding Programs

Peng-Yeng Yin, National Chi Nan University, Taiwan
Shyong Jian Shyu, Ming Chuan University, Taiwan
Shih-Ren Yang, Ming Chuan University, Taiwan
Yu-Chung Chang, Ming Chuan University, Taiwan

ABSTRACT

Due to the explosive and growing size of the genome database, the discovery of gene has become one of the most computationally intensive tasks in bioinformatics. Many such systems have been developed to find genes; however, there is still some room to improve the prediction accuracy. This paper proposes a reinforcement learning model for a combination of gene predictions from existing gene-finding programs. The model learns the optimal policy for accepting the best predictions. The fitness of a policy is reinforced if the selected prediction at a nucleotide site correctly corresponds to the true annotation. The model searches for the optimal policy which maximizes the expected prediction accuracy over all nucleotide sites in the sequences. The experimental results demonstrate that the proposed model yields higher prediction accuracy than that obtained by the single best program.

Keywords: Bioinformatics, DNA Sequence, Gene-Finding, Genome Database, Reinforcement Learning

1. INTRODUCTION

There is an explosive growth in the amount of sequenced nucleotides from a number of genome projects. Several million bases of genomic DNA are sequenced daily and made available to the public. The genome annotation, one of the most important works of genome project, is to find all existing genes on the genomic DNA sequences. Conventionally, the genome annotation can be divided into three steps, namely automatic annotation, manual annotation and experimental verification. The automatic annotation is the main task of bioinformatics. This is because manually annotating the coding regions of genes on all genomic sequences from scratch is impractical; instead, the sequences...
should be pre-analyzed in a high-throughput computational way.

To date, many gene-finding programs have been developed to annotate the newly sequenced genomes. These programs can be broadly classified into two classes. Homology-based methods discover genes using a systematic database search. They are based on the fact that homologous genes may have similar gene structures (Kulp et al., 1997; Korf et al., 2001), such as AAT (Huang, 1997), GeneBuilder (Milanesi et al., 1999) and CRASA (Chuang et al., 2003). However, researchers have found that only about 50% of the annotated genes have similar homologues in the database (Uberbacher et al., 1996; Dunham et al., 1999). This is due to the amount of known proteins are currently limited.

The second class of gene-finding programs consists of ab initio methods which are based on machine learning techniques such as artificial neural networks (GRAIL) (Xu et al., 1997), GeneParser (Snyder & Stormo, 1995), discriminant analysis (GeneFinder) (Solovyev et al., 1994), MZEF (Zhang, 1997), hidden Markov models (Genie) (Kulp et al., 1996), GENSCAN (Burge & Karlin, 1997), and HMMgene (Krogh, 1997). Although these gene-finding programs have been demonstrated with high prediction accuracy in certain species, we still lack a universal program that can report satisfactory accuracy in general cases. Most researchers attempted to develop a new gene-finding program that can attain better prediction accuracy than existing ones; however, they overlook the fact that different gene-finding methods may be considerably complementary to one another. The fusion of multiple experts has been empirically shown to be superior to using single expert alone in many machine learning applications (Richardson & Domingos, 2003; Yin et al., 2005). We argue that even a worse gene-finding program can correct part of the predictions produced by a novel gene predictor. For instance, a 47,338 bp of human DNA is annotated for determining the gene structures (that is to point out where is exon or intron), GENSCAN can correctly identify 47,083 bp. For the remaining 255 bp that are not well annotated, all of them can be identified either via AAT (128/255 bp) or using HMMgene (156/255 bp). It reveals a chance for synergy among variable programs. Therefore, the combination of multiple gene-finding programs may get better gene annotation rather than using any single best program.

So far some combinations of gene-finding programs for improving gene annotation were reported, such as GeneNomi (Harris, 1997) that combined BLAST, GRAIL, GeneFinder, and Genie to make better predictions, however, the details for combining these tools were not described. Murakami and Takagi (1998) also analyzed several combination methods, namely AND, OR, HIGHEST, RULE, and BOUNDARY, with four programs, that is FEXH, GeneParser3, GENSCAN, and GRAIL2. The method AND labels the exon candidates as the intersection of the predicted exon regions by the programs, while the method OR determines the exon regions as the union of the program-predicted exons. As for the method HIGHEST, the exon candidates are those regions which have the highest score among the programs. The method RULE determines the predictions in accordance with a priority order of the programs based on a previous empirical study. Finally, the method BOUNDARY determines the best coding-noncoding boundary given the score and boundary type by the programs. The five combination methods are simple and ad-hoc, they cannot accommodate the correlations among programs and the dependence between any two adjacent nucleotides. For instance, given the predictions (coding or non-coding) by three programs at any two adjacent nucleotides there are totally 64 possible combinations, however, most of the combinations are not differentiated by those methods. In addition, Rogic et al. (2002) also proposed three methods for combining predictions with GENSCAN and HMMgene. They focused on improving exon level accuracy by union or intersection of predicted exon regions considering probabilistic scores and reading frame consistency. The accuracy improvement on a newly assembled dataset is 7.9% over the best of individual programs. These methods are also
12 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage: www.igi-global.com/article/reinforcement-learning-improving-gene-identification/64643?camid=4v1

This title is available in InfoSci-Journals, InfoSci-Journal Disciplines Computer Science, Security, and Information Technology. Recommend this product to your librarian: www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

Evolutionary Approaches for the Extraction of Classification Rules
www.igi-global.com/article/evolutionary-approaches-for-the-extraction-of-classification-rules/113826?camid=4v1a

A Local Search Approach to Solve a Financial Portfolio Design Problem
www.igi-global.com/article/a-local-search-approach-to-solve-a-financial-portfolio-design-problem/125863?camid=4v1a

A Hybrid Genetic Algorithm for Optimization of Two-Dimensional Cutting-Stock Problem
www.igi-global.com/chapter/hybrid-genetic-algorithm-optimization-two/63804?camid=4v1a
Performance Enhancement of Differential Evolution by Incorporating Lévy Flight and Chaotic Sequence for the Cases of Satellite Images
