Exon prediction problem using genetic algorithm as an approach for hypothesis testing

Edgar Augusto G. G. do Amaral
Federal University of Rio de Janeiro – PESC/COPPE/UFRJ
21941-972, Tecnology Center, RJ
E-mail: edgar@cos.ufrj.br

Rubem P. Mondaini
Federal University of Rio de Janeiro – PESC/COPPE/UFRJ
21941-972, Tecnology Center, RJ
E-mail: mondaini@cos.ufrj.br

ABSTRACT

The gene identification problem can be formulated as the deduction of the amino acid sequences encoded in a given DNA genomic sequence [1]. This is an important but difficult problem, especially in eukaryotes, where genes are often split into exons separated by introns. The beginning or end of these fragments is specified in the genome sequence by a few types of signals encoded in the primary DNA sequence of a gene.

Using currently available detection methods, it is usually impossible to distinguish signals truly processed by the cellular machinery from those actually not functional. In order to predict gene structure by processing only DNA sequence signals often results in a computationally intractable combinatorial explosion of potential products. As a result, any gene prediction method that relies on these signals has to be able to distinguish between false exons and true exons.

The main objective of this work is to derive and compare methods to improve the performance of some gene prediction methods, such as Geneid [2], for the new proposed method to be applied to the genetic algorithm architecture. These methods view the exon prediction step as a problem of hypothesis testing. This leads to the possibility of introducing a probabilistic scoring system based on P-values, which could be used as an additional input in the dynamic programming algorithms used to build the final gene prediction from the candidate exons. Other dynamic algorithms, such as genetic algorithms, could also be implemented using Pvalues.

This work considers three different ways of deciding whether, given an exon this value should be considered statistically as greater than zero i.e. significant: *Sum-of-scores tests*, *Intersection-union tests* e *Meta-analysis based tests*. All three methods require the sampling distribution of the test statistic under the null hypothesis to be known. This has been approximated using the approaches: *Resampling* e Monte Carlo.

This work proposes a new method for detecting coding sequences. The proposal is to extend using with genetic algorithms to detect alternative splicing or to compare genomes.

References

[1] R. Guigó et al. An assessment of gene prediction accuracy in large DNA sequences. Genome Res., 10, 1631–1642, 2000.

[2] G. Parra et al. Geneid in Drosophila. Genome Res., 10, 511–515, 2000.