Software for annotation of protein coding genes in yeast mitochondrial genomes

Juraj Mešťánek
UK, Bratislava, Fakulta matematiky, fyziky a informatiky

Proteins are encoded in DNA by segments called genes. A gene generally consists of two types of segments: exons and introns. Introns are non-coding parts of genes that need to be removed before the process of translation to protein can start. Coding segments, that are left after the removal of introns, are called exons. The problem of gene finding is to identify exons and introns in a DNA sequence. Most of the research in gene finding concentrates on genes in nuclear genomes and there are many programs that address this problem. In this paper we present a software tool for automated computational prediction of protein coding genes in yeast mitochondrial genomes. Yeast mitochondrial genes lack the clear exon boundary rules typical for nuclear genes. This makes identifying precise exon boundaries much harder. On the other hand, mitochondrial genomes are short and contain only a small set of well-conserved genes which allows us to use strategies that are more difficult to apply on nuclear genes.

Our tool is based on conditional random fields (CRFs). CRFs allow us to incorporate information from many information sources, even if it does not have a probabilistic interpretation. This would be problematic in a more traditional hidden Markov model approach.

To produce accurate annotation, our tool combines information from several different sources. We use Exonerate to align reference proteins extracted from model organisms to the genome being annotated. Genes coding these proteins are very well conserved across the studied organisms so the resulting alignment gives a very good approximation of exon and intron positions. We also use RNAWeasel to predict the positions of introns based on their characteristic structural motifs. Finally, we use multiple alignment of mitochondrial genomic sequences from several yeast species to look for evolutionary signatures typical for protein-coding regions. These three sources of information as well as the studied nucleotide sequence form a set of observations used in our CRF model to predict positions of exons and introns.

We have tested our tool on genes from 33 mitochondrial genomes. Currently, we predict 78% of genes and 70% of exons perfectly. We are working to further improve prediction accuracy of our tool and to make it available and easy to use for the life science community. The paper is meant to be the author's master's thesis.