CodAn: predictive models for the characterization of mRNA Transcripts

Pedro Gabriel Nachtigall, Andre Y. Kashiwabara, Alan Durham

Laboratório Especial de Toxinologia Aplicada (LETA), Instituto Butantan, São Paulo, Brazil

Abstract

The complete characterization of the coding sequences (CDSs) and untranslated regions (UTRs) of transcripts is an essential step on transcriptome annotation and expression profile analysis. First, it defines which proteins should be synthesized by the messenger RNAs and are part of the proteome of the organism. The incorrect characterization of CDSs can lead to the prediction of non-existent proteins. Wrong protein predictions can eventually compromise knowledge if annotation databases are populated with similar incorrect predictions made in different genomes. Also, the correct identification of CDSs is important for the characterization of the UTR landscape, whereas the 3'UTR and 5'UTR are known as important regulators of the mRNA fate and translate process. Here, we present CodAn, a new computational approach to predict CDS and UTR sequences directly from transcriptome sequences of any Eukaryote species, such as RNAseq assembly data. CodAn can be applied to full or partial transcripts and presents a better performance predicting the whole CDS than other approaches. CodAn requires low computational resources and can be used on any standard desktop computers, and, for large jobs, can use the parallel processing capabilities of large multi-core servers. The data generated by CodAn can be used to improve genome annotation and help further experiments focused on understanding the evolution and biology of CDS and UTR sequences.

Funding: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 - Process number 88887.177457/2018-00.