DOI 10.18699/BGRSSB-2018-162

Algorithms for prediction and analysis of regulatory regions

T. Tatarinova[1]*, V. Solovyev[2], N. Alexandrov[3], A. Baranova[4], A. Kel[5, 6]
[1] *University of La Verne, La Verne, CA, USA*
[2] *Softberry, Inc. Mount Kisco, NY, USA*
[3] *Inari, Inc, Boston, MA, USA*
[4] *George Mason University, VA, USA*
[5] *geneXplain GmbH, Wolfenbuettel, Germany*
[6] *Institute of Chemical Biology and Fundamental Medicine, Novosibirsk, Russia*
** e-mail: ttatarinova@laverne.edu*

Key words: promoter, deep learning

Motivation and Aim: They say that the chain is only as good as its weakest link. Current methods of genome annotation are capable of accurate prediction of coding regions, but are failing in promoter prediction. Accurate identification of transcription start sites and core promoter regions remains an unsolved problem.
Methods and Algorithms: We will present a comprehensive analysis of genomic features associated with promoters in several plant genomes, and demonstrate how probabilistic integrative algorithms succeed in accurate prediction of transcription start sites. We developed models using distributions of sequence polymorphisms, RNA sequencing reads on genomic DNA, methylated nucleotides, transcription factor binding sites, as well as relative frequencies of nucleotides and their combinations.
Results: Accuracies of promoter-prediction methods differ between species and functional classes of genes, and we will present an approach to select the optimal method for promoter prediction for a studied genome. We have identified three distinct classes of TFBS that show different positional preference with respect to TFBS. We have also demonstrated evolutionary conservation of distribution of TFBS between plant species.