ISSN 2334-6590
UDK 57 (051)

Biologia Serbica

Book of Abstracts
Belgrade BioInformatics Conference 2018

Volume 40 - No. 1
(Special Edition)
2018

BELBi 2018

ИМГГИ
IMGGE

Department of Biology and Ecology
Faculty of Sciences
University of Novi Sad

ragp: An R toolbox for mining plant Hydroxyproline rich glycoproteins

Danijela Paunović[1], Milica Bogdanović[1], Slađana Todorović[1], Ana Simonović[1] and Milan Dragićević[1*]

[1]Institute for Biological Research "Siniša Stanković", University of Belgrade, Department of Plant Physiology, Bul. despota Stefana 142, 11000 Belgrade, Serbia

Abstract:
Plant Hydroxyproline rich glycoproteins (HGRPs) comprise a highly diverse family of cell wall macromolecules, involved in a wide array of physiological functions such as cell expansion, somatic embryogenesis, self-incompatibility, signaling and pathogen responses. Due to biased amino acid composition, abundant in disorder promoting residues, HRGPs are intrinsically disordered proteins. The lack of a stable structure lessens the sequence constraints imposed on these proteins and hampers efforts for homology based identification. Current mining approaches, based on identifying sequences with characteristic motifs and biased amino acid composition, are limited to prototypical sequences.

Herby we present ragp, an R package for HGRP mining with a pipeline which emphasizes finding chimeric and short HRGP's which is especially useful for identification of arabinogalactan proteins. The ragp pipeline exploits one of HGRP key features, the presence of hydroxyprolines which represent glycosylation sites. These sites are identified using a gradient boosting model trained on plant sequences with experimentally determined hydroxyprolines, based on the local (21-mer) sequence around the target prolines. The model was validated on a set of sequences which were not used during the model building, as well as by using several resampling approaches. Apart from prediction of proline hydroxylation main ragp features include efficient communication with web servers for prediction of N-terminal signal peptides and GPI modification sites, sequence annotation by querying hmmscan, GO enrichment based on predicted pfam domains, and the ability to classify prototypical HRGPs.

ragp represents the first implementation of a HRGP mining workflow in the R statistical language. It implements common strategies for finding and classifying HRGP sequences along with an optional step where proline hydroxylation is estimated which leads to increased sensitivity and specificity of the filtered sequences. Since R is one of the leading bioinformatics platforms, the filtered sequences can be further analyzed by many specialized packages using the same environment.

Keywords:
bioinformatics, data mining, gradient boosting, hydroxyproline rich glycoproteins, arabinogalactan proteins

This work has been funded by MESTD, Republic of Serbia (projects TR31019, OI173024).

*Corresponding author, e-mail: mdragicevic@ibiss.bg.ac.rs