LncRNAplant-Finder: a tool for predicion of long non coding RNAs in plants

Tatianne da Costa Negri[1], Pedro Henrique Bugatti[1], Priscila Tiemi Maeda Saito[1], Douglas Silva Domingues[1,2] e Alexandre Rossi Paschoal[1]

[1]Programa de Pós-Graduação em Bioinformática - PPGBIOINFO, Universidade Tecnológica Federal do Paraná, Campus Cornélio Procópio; [2]Departamento de Botânica, Instituto de Biociências, Universidade Estadual Paulista, Campus de Rio Claro
** Corresponding author: paschoal@utfpr.edu.br*

Long non coding RNAs (LncRNAs) correspond to a eukaryotic non-coding RNA class with more than 200 nucleotides in lenght. They have emerging attention in the last years as a potential layer of gene expression in cells. However, lncRNAs mechanisms in plants are still poorly known. Moreover, there is a lack of specific computational approaches for lncRNA prediction in plants, considering that the biological mechanism of this ncRNA class is different from mammals, which there are several tools for prediction. Having this in mind, we present the LncRNAplant-Finder, an approach for lncRNA identification in plants. To built this tool, we used publicly avaliable lncRNA and transcript (mRNA) sequences from six plant genomes: *Arabidopsis thaliana, Cucumis sativus, Glycine max, Oryza sativa, Populus trichocarpa* and *Setaria italica*. All the data was extracted from the public databases PLNlncRbase, GREENC and Phytozome, where we used 22,543 lncRNAs and 29,960 transcripts. We applied pattern recognition techniques in a total of 85 features based on sequence and structure from lncRNAs and transcripts (e.g. GC content, ORF, dinucleotide and trinucleotide distribution) in order to select the best features for classification. Sequences were also processed using: (i)- CD-Hit-EST: to avoid sequence redundancy; (ii)- txCDSPredict: for ORF prediction; (iii)- in-house PERL scripts to calculate di and tri-nucleotides frequency, GC context, normalization and generating an ARFF file. All feature selection and classification processes were done using Weka 3.8.0. We detected 16 best features for classification after feature the selection process. These features were used to compare six classification methods. The J48 method obtained the best results with: (i)- Correctly Classified Instances (CCI), ≃97%; (ii)- Incorrectly Classified Instances (ICI), ≃ 3%; (iii)- Correct lncRNAs (CL), 22.021 (≃97,7%); (iv)- Correct Transcripts (CT), 28.812 (≃96,25%); (v)- Error lncRNA (EL), 522 (≃2,3%); (vi)- Error Transcripts (ET), 1.148 (≃3,9%). These results point out a promissing approach to help lncRNA identification in plant genomes.

Support: CNPq (#454505/2014-0); Fundação Araucária: Nº 019/2015.