Proposal of a data mining pipeline to improve bacterial small RNA prediction

Fabio Reinoso Vilca, Sabrina de Azevedo Silveira, Fabio Ribeiro Cerqueira

Departamento de Informática, Universidade Federal de Viçosa, Viçosa-MG

In the last years, the discovering of novel bacterial small RNAs (sRNAs) became relevant due to their essential roles in important cellular activities. Such molecules are key for a number of mechanisms such as: Regulation of outer membrane protein expression, iron homeostasis, quorum sensing, and bacterial virulence. Prediction of sRNA is a challenging issue in bioinformatics, i.e., the current computational tools have low precision and sensitivity. However, the development of predictive methods are of fundamental importance to narrow the number of costly and time-consuming sequence validations on the laboratory workbench. In this work, we extract different kind of features of a putative sRNA sequence to perform the prediction task. Some important features are: Sequence-based features, secondary structure features, base-pair features, triplet sequence structure, and structural robustness. Additionally, we apply the InformationGain feature selection algorithm to select the best 50 features over the initial 243 set of features. Our preliminary results show that the most relevant features are those related to the secondary structure. Our dataset is composed of known and experimental-validated sRNAs obtained from the BSRD database, as well as different kind of non-sRNA sequences such as: Shuffled sequences generated from real sRNAs, coding sequences, and other noncoding types of RNAs (tRNAs, rRNAs) from 5 different bacteria substrains pertaining to 4 different families, including one cyanobacteria. To handle the imbalance problem of our dataset, which contains 824 sRNAs and 3032 non-sRNAs sequences, we use the SMOTE technique. Finally, we apply the random forest learning algorithm for the classification task. The current results of our approach reached an accuracy of 92.66%, an specificity of 92.66%, and an sensitivity of 92.7%.

This research is supported by CAPES, CNPq, and FAPEMIG