Classification of Transposable Elements through Convolutional Neural Networks

Murilo Horacio Pereira da Cruz, Douglas Silva Domingues, Priscila T M Saito, Alexandre R Paschoal, Pedro Henrique Bugatti

São Paulo State University

Abstract

Transposable Elements (TEs) are the most represented sequences occurring in eukaryotes. They can change their location and generate multiple copies of themselves throughout genomes. This action can cause significant effects in organisms, such as the regulation of gene expression. There are several types of these elements which are classified into two classes, nine orders, and 29 superfamilies. Sequences are classified in a hierarchic way, in which classes are divided into orders and orders into superfamilies. The correct classification of these sequences is still a challenging quest. Due to the rapid increase in the number of sequenced genomes, the manual classification of these sequences is no longer feasible. Besides, automatic methods are mostly based on sequence alignment, a strategy that demands high computational costs. Therefore, novel strategies are required for this problem. To fill this gap, we present an automatic TE classification approach through a Convolutional Neural Network (CNN). Unlike traditional machine learning algorithms, that use handcrafted features to classify data, CNN can learn the best representation for the data and how to correctly classify it, given it is a representation learning algorithm. Few methods on the literature provide the classification of these sequences into the superfamily level. Superfamilies of the same order tend to share similar structures, i.e. the type of repeats, sequence length, and protein domains, thus providing a challenge to handcrafted features based methods. We evaluate the performance of CNNs on the classification of three datasets built using sequences from seven databases. We compared our results to TEclass, a method that uses Support Vector Machines to classify TEs into one class and three orders. Our approach obtained an accuracy of 92.3% on the classification of RepBase sequences from nine superfamilies and 94.6% on the classification of sequences from all seven databases into 10 superfamilies. We also obtained 98.1% on the classification of three RepBase orders and 94.3% on the classification of sequences from all databases from four different orders.

Funding: This work is supported by CAPES