Using CORAZON to investigate functionally and evolutionarily related coding and non-coding transcripts

Thaís de Almeida Ratis Ramos[1], Thaís Gaudencio[2], Vinicius Maracaja Coutinho[3], José Miguel Ortega[4]

1 UFRN
2 UFPB
3 UNIVERSIDAD MAYOR
4 UFMG. LABORATÓRIO DE BIODADOS.

Abstract

Machine learning is a subfield of computer science that developed from the study of pattern recognition and computational learning theories in artificial intelligence. These methods operate through the construction of a model based on the set of inputs, in order to make data predictions. Due to the large quantity of biological data generated in large-scale genomics and transcriptomics projects, an intense demand to use techniques provided by artificial intelligence, the usage of tools based on machine learning methods became widely used in bioinformatics. Unsupervised learning is the machine learning task of inferring a function to describe the hidden structure from unlabeled data. The inductor analyzes the examples provided and tries to determine if some of them can be grouped in any way, forming clusters. Here we developed an online tool calling CORAZON (Correlation Analyses Zipper Online) that include 3 unsupervised machine learning algorithms: Mean shift, K-means and Hierarchical with the bioinformatics purpose. Furthermore, we implemented 4 normalization methodologies: Transcripts Per Kilobase Million (TPM), base-2 log, instance normalization and normalization by the highest attribute value for each instance. Moreover, the user has a option to cluster the data removing each attribute to see the results, to observe the attributes influence. Here, we used our tool to study the coding and non-coding genes from Uhlen, Fantom and Encode databases. We found clusters well defined with genes of each of the two classes and analyzed biological processes determined by GO Enrichment Analysis and gene ages defined by Life Cycle Assessment (LCA). Normally, clusters with more codings are associated with cellular, metabolics, transports and systems development processes. Clusters with more non-codings are involved with detection of stimulus, sensory perception, immunological system, and digestion. We also observed that clusters with more than 80% of non-codings, more than 40% of their coding genes are recents appearing in mammalian class and the minority are from eukaryota class. Otherwise, clusters with more than 90% of coding genes, have more than 40% of them appeared in eukaryota and the minority from mammalian. Clusters without these criterias, have the majority of their coding genes arise from Eumetazoa. Therefore, the CORAZON tool can help in the large quantities analysis of genomic data, facilitating to comprehend the relations between these instances. In addition, as future work, it will be possible to understand the evolutionary history of these sequences and the associated transcription factors.

Funding: UFRN