Using machine learning to cluster genes and tissues according to their functions through gene co-expression

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

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

Abstract

The creation of gene expression encyclopedias possibilities the understanding of genes groups that are co-expressed in differents tissues and comprehend gene clusters according to their functions. The advent of machine learning, with unsupervised methods without needing to define the number of clusters a priori on the clustering process, is possible to map large data sets. This would be the first step to understanding the performance of transcription factors in the regulation processes of gene expression. The purpose of this work is to evaluate genes coexpression by tissue and function through gene expression data. For that, were tested 3 databases: Uhlen, Fantom and Encode. As pre-processing data, four normalization types were tested and adopted the combination of two of them: Transcripts Per Kilobase Million (TPM) and base-2 log. In the clustering process we use 2 machine learning algorithms: K-means and Hierarchical implemented in an online tool calling CORAZON (Correlation Analyses Zipper Online). To select the best number of clusters were used: Bayesian information criterion (BIC) followed by the derivative of the discrete function and Silhouette. Furthermore, in the Hierarchical we test eight linkage criterions and adopted the Ward's method. The first database with 32 tissues had an optimal number of clusters equal to 9, the second with 56 tissues, 11 clusters and the last with 13 tissues, 7 clusters. We observed that hierarchical method and K-means generated exactly the same clusters, only a few had some slight variation among their components. However, we can observe groups related to glands, cardiac tissues, muscular tissues, tissues related to the reproductive system and in all three groups are observed with a single tissue, such as testis, brain and bone-narrow. The same uniformity behavior was found in the functional analysis of the gene groups after clustering. In the first database were analyzed 44594 genes in 9 clusters; 21080 genes in the second database grouped into 11 clusters and the third database grouped 42355 genes into 10 clusters. In relation to the genes clusters, we obtained several clusters that have specificities in their functions: detection of stimulus involved in sensory perception, reproduction, synaptic signaling, nervous system, immunological system, system development, and metabolics. These results are preliminary but show the methods potential and the possibilities of analyzing transcription factors in the regulation process of these genes, making possible the evolutionary history study of genes and the biological system regulatory map.

Funding: UFRN