EntropyClusterGenes: a R package for clustering genes according ontologies and pathways

André Luiz Molan[1], Carlos Biagi Jr[2], Giordano Bruno[1], Jose Rybarczyk-Filho[1],

1 UNESP
2 UNESP - Botucatu/SP

Abstract

NGS technologies have transformed the way we study living organisms. By the application of different techniques, such as RNA-seq, numerous species can be studied at a relatively low cost. The amount of data generated, however, is huge. It's not so easy to analyze it, demanding computational tools increasingly efficient and with different approaches, highlighting those with a focus on functional analysis. In this way, we developed EntropyClusterGenes, a R package capable of clustering genes according to their respective Gene Ontology (biological processes - BP, molecular functions - MF, cellular components - CC and KEGG pathways) and determining the significance of such sets based on the expression values ??of their genes. We start with a text file containing a gene list and their expression values in two comparative samples (e.g., experiment and control). Through the clusterProfiler and topGO R packages, linking them to online databases of different species, we group the genes according to their respective functions. The behavior of the genes within each group is shown by the relative calculation of two variables: gene activity and gene diversity. The first one characterizes the set just according to the expressed value of the genes contained therein, while the second one measures the gene diversity within the set by the use of a normalized Shannon's entropy function. Each set is assigned with a p-value, obtained through the bootstrapping statistical method and multiple comparisons between the activity and gene diversity variables. To determine which sets are significant, we apply the FDR (False Discovery Rate) statistical method, considering, by default, values ??lower than 0.05. To demonstrate the use of the EntropyClusterGenes, we applied the tool in two data sets of microarray (Homo sapiens and Rattus norvegicus) and a RNA-seq data set of Aedes aegypti. The number of significant groups found varied according to the species, sample and function studied. We considered a FDR of 0.05 for all species. Despite the number of bootstrapping steps, we used 10,000 for microarray and 500,000 for RNA-seq. In the case of H. sapiens, from a group of 20502 genes, obtained by comparing different doses of etoposide in the liver, for BP, CC, MF and KEGG, were identified, respectively, 5825, 696, 10082 e 305 groups, with an significance average percentage per Gene Ontoloy of 8.52%.

Funding: CNPq processes 473789/2013-2 and 134469/2016-0.