TITLE: CLUSTREFGENES: A WEB TOOL TO IDENTIFY CANDIDATE REFERENCE GENES BASED ON RNA-SEQ DATA

AUTHORS : FRANCO, E. F.; MAUES, D.; PINHEIRO, K.C.; RONNIE, A.; GUIMARAES, L.; AZEVEDO, V.; SILVA, A.; GOSH, P.; MORAIS, J.; RAMOS, R. T. J.

INSTITUTIONS: 1-INSTITUTE OF BIOLOGICAL SCIENCES, LABORATORY OF BIOLOGICAL ENGINEERING, FEDERAL UNIVERSITY OF PARA (R. AUGUSTO CORRÊA, 1 - GUAMÁ, BELÉM - PA, 66075-110)-BRAZIL; 2-DEPARTMENT OF COMPUTER SCIENCE,COMPUTER SCIENCE POSTGRADUATE PROGRAM (PPGCC), FEDERAL UNIVERSITY OF PARA (BELEM,PARA)-BRAZIL; 3-VALE TECHNOLOGY INSTITUTE (BELEM, PARA)- BRAZIL;4-INSTITUTE OF BIOLOGICAL SCIENCES, FEDERAL UNIVERSITY OF MINAS GERAIS-UFMG (BELO HORIZONTE, MINAS GERAIS)-BRAZIL. 5-DEPARTMENT OF COMPUTER SCIENCE, VIRGINIA COMMONWEALTH UNIVERSITY (RICHMOND, VA)-USA

ABSTRACT:
Reference Genes (RG) or housekeeping (HKG) are constitutive genes required for the maintenance of basic cellular functions. Thus, RGs are expressed in all cells of an organism under both normal and pathophysiological conditions. However, some RGs are expressed at relatively constant rates in most non-pathological situations, although in recent times different studies have reported variations in the reference genes under different experimental treatments, time variations or cell types. Hence, the expression of RGs is used as a reference point in the expression levels of other genes in analyzing gene expression. RNA-Seq is a high-throughput method that allows the measurement of gene expression profiles in a target tissue or an isolated cell, that gene expression data is essential to understand many biological processes in different organisms in relation to its environment. The machine learning (ML) techniques, which enable the collection and identification of valid, new, potentially usable and understandable patterns and knowledge based on a dataset, which can lead to the detection and identification of possible RG candidates. ML methods have also been applied in different genomic areas to enable the interpretation of large datasets, including those related to gene expression. Within machine learning methods, clustering algorithms are based on unsupervised learning and are used to cluster objects based on the intrinsic information contained in the data and their relationships. This study reports ClustREFGenes a web tool to identify reference genes candidates in silico through clustering techniques using RNA-seq data. The method used validated reference genes set to identify new references genes candidates based on Euclidean distance. The tool is developed in Python and R as back-end programming languages. PHP, JavaScript and Flask are used as the front-end programming language. ClustREFGenes was developed to be executed in any modern internet browser. Also, it has a clean and easy-to-use graphic interface. It's not required any kind of installation of any tool or software and users can execute projects without previous registration. This tool has enabled the identification of stable RG candidates in RNA-seq data from Corynebacterium pseudotuberculosis and Escherichia coli strains. These genes showed stable expression under different stress conditions as well as low variation index and fold changes. Furthermore, some of these genes were already reported in the literature as RGs or RG candidates for the same or other bacterial organisms, which reinforced the accuracy of the proposed method.

Keywords: genomics**,** housekeeping, machine learning, RNA-seq

Development Agency: CNPq, CAPES