Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

An algorithm to infer similarity among cell types and organisms by examining the most expressed sequences

S.A.P. Pinto[1] and J.M. Ortega[2]

[1]Departamento de Bioquímica e Imunologia, Laboratório de Biodados,
Instituto de Informática/Barreiro, Pontifícia Universidade Católica de
Minas Gerais, Instituto de Ciências Biológicas,
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[2]Departamento de Bioquímica e Imunologia, Laboratório de Biodados,
Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais,
Belo Horizonte, MG, Brasil

Corresponding author: J.M. Ortega
E-mail: miguel@icb.ufmg.br

Genet. Mol. Res. 7 (3): 933-947 (2008)
Received June 2, 2008
Accepted August 11, 2008
Published September 30, 2008

ABSTRACT. Following sequence alignment, clustering algorithms are among the most utilized techniques in gene expression data analysis. Clustering gene expression patterns allows researchers to determine which gene expression patterns are alike and most likely to participate in the same biological process being investigated. Gene expression data also allow the clustering of whole samples of data, which makes it possible to find which samples are similar and, consequently, which sampled biological conditions are alike. Here, a novel similarity measure calculation and the resulting rank-based clustering algorithm are presented. The clustering was applied in 418 gene expression samples from 13 data series spanning three model organisms: *Homo sapiens*, *Mus musculus*, and *Arabidopsis thaliana*. The initial results are striking: more than 91% of the samples were clustered as expected. The MESs (most expressed sequences) approach outperformed some of the most used clustering algorithms applied to this kind of data such as hierarchical clustering and K-means. The clustering performance suggests that the new similarity measure is an alternative to the traditional correlation/distance measures typically used in clustering algorithms.

Key words: Gene expression; Clustering; Similarity measure;
Most expressed sequences