Identification of candidate genes associated to several cell lines of the cancers tumors using text-mining, co-expression network analysis and comparative genomics: A STATIS-ACT methodology approach

M L Zingaretti[1], *Johanna Demey-Zambrano*[23], *Grace Dean*[3], *O Ruiz*[4], *Julio Di Rienzo*[5], *M P Galindo-Villardón*[2], *J L Vicente-Villardón*[2], *J R Demey*[764]

[1] Universidad Nacional de Villa María. Provincia de Córdoba, Argentina. [2] Departamento de Estadística. Universidad de Salamanca. Salamanca, España. [3] School of Nursing, University at Buffalo. Buffalo, NY, USA. [4] Escuela Politécnica Superior del Litoral. Guayaquil, Ecuador. [5] Universidad Nacional de Córdoba. Córdoba, Argentina. [6] Instituto de Estudios Avanzados. Caracas, Venezuela. [7] Proyecto Prometeo. SENESCYT, Ecuador.

Correspondence author: jdemey@idea.gob.ve

In the last years, the data of microarrays has not only gained a great importance but also it is availability for the public has increase. The "omics" technologies allow quantitative knowledge of hundreds of biological data of complex nature and have enabled the opportunity of study simultaneously, based on multiple datasets, the expression levels of thousands of genes over the effects of certain treatments or diseases. However, the joint analysis of the different subspaces that generate these technologies and their relations is not simple. Several statistical methods have been developed to handle these problems and to calculate a consensus from data matrices. STATIS-ACT is one of the families of methods that are concerned with analysis of data arising from several configurations and is a powerful technique to compare subspaces. The aim of this paper is to combine STATIS-ACT and Biplot methodology to study the relationships between genes expressions of multiple microarrays platforms from NCI-60 database. We found the Euclidean image of the compromise generated by the STATIS-ACT methodology from scalar products of the individual studies indicated that the cell- lines originating from the colon, leukemia, melanoma and CNS were grouped. In contrast, the breast cancer cell line present high heterogeneity. The projected genes in compromise were grouped into six groups using cluster algorithm and the relationship was established between groups of genes and of cell lines types. The genes with strongest association to a cell line were those projected in the same direction and they were over-expressed in these cell lines. Finally, text mining and network building tools were used to establishing relationships between selected genes and those reported in the literature associated with cancer. The results obtained confirmed the capabilities of the proposed methodology to selected genes, detect the best classification and to the study relationships between cell lines tissue and gene expression in transcriptomes studies.

Keywords: Biplot, Candidate genes, Co-expression, Microarrays, NCI-60 database, Omics, STATIS-ACT methodology.

AMS: 62H30