In silico markers: an evolutionary and statistical approach to select informative genes of human breast cancer subtypes.

Shib Sankar Bhowmick, Debotosh Bhattacharjee, Luis Rato

Abstract

Background Recent advancement in bioinformatics ofers the ability to identify informative genes from high dimensional gene expression data. Selection of informative genes from these large datasets has emerged as an issue of major concern among researchers. Objective Gene functionality and regulatory mechanisms can be understood through the analysis of these gene expression data. Here, we present a computational method to identify informative genes for breast cancer subtypes such as Basal, human epidermal growth factor receptor 2 (Her2), luminal A (LumA), and luminal B (LumB). Methods The proposed In Silico Markers method is a wrapper feature selection method based on Least Absolute Shrinkage and Selection Operator (LASSO), Covariance Matrix Adaptation Evolution Strategy (CMA-ES) and Support Vector Machine (SVM) as a classifer. Moreover, the composite measure consisting of relevance, redundancy, and rank score of frequently appeared genes are used to select informative genes. Results The informative genes are validated by statistical and biologically relevant criteria. For a comparative evaluation of the proposed approach, biological similarity score designed on semantic similarity measure of GO terms are investigated. Further, the proposed technique is evaluated with 7 existing gene selection techniques using two-class annotated breast cancer subtype datasets. Conclusion The utilization of this method can bring about the discovery of informative genes. Furthermore, under multiple criteria decision-making set-up, informative genes selected by the In Silico Markers are found to be admirable than the compared methods selected genes.

Keywords: Breast cancer subtype· Biological analysis· Gene selection· Messenger RNA· Statistical analysis

https://doi.org/10.1007/s13258-019-00816-8