Prediction and evaluation of driver genes based on TCGA somatic mutations data in breast invasive carcinoma patients

L. Mirsadeghi[a,b]*, E. Mirzaei[C], K. Kavousi[b], A. Banaei-Moghaddam[b], R. Hajihosseini[a]
[a] Payame Noor University, Tehran, Iran
[b] Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[c] MsC graduate of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
*l.mirsadeghi@gmail.com

Abstract: Every day, our body cells influenced by mutations that induce small changes to the genetic material including genes. These genes drive some diseases including complex diseases. Among the masses of our genes, these defective genes are called "driver genes". In complex diseases such as cancers, there are anomalies in more than a single gene. Breast invasive carcinoma as a complex disease is the most common cancer in women worldwide. This metastatic disease has now become the most frequent cancer site among the urban women in Iran [1]. So, researchers are investigating the possibility of using bioinformatics tools to help detect and prioritize these driver genes about this cancer [2]. In this research, we have applied a bioinformatics software tool, named MutSigCV (Mutation Significance (CV= covariates);) that has been developed by Broad Institute. MutSigCV v1.4 bases on “Mutation Significance”. It accepts whole genome or whole exome sequencing data from multiple samples, with information about point mutations and small insertions and deletions (INDELs). Then it analyzes lists of discovered mutations, and identifies genes that are mutated more than one would expect by chance [3]. At this study and for applying MutSigCV v1.4, we used somatic mutations data and downloaded five files (.maf), from the Cancer Genome Atlas (TCGA) data set which are involved four open access data files and one controlled data access file. The controlled data file includes 47,657 somatic mutations identified in 16,378 genes from 695 breast invasive carcinoma patients. After using MutSigCV, we firstly compared outputs related to opened data access files. The result of comparison between predicted driver genes (P-value<0.05) showed that six genes including of *TP53*, *PTEN*, *PIK3CA*, *RB1*, *ZFP36L1* and *NF1* are overlapping. It is interesting that these driver genes have also been predicted in controlled data access file output with P-value less than 0.004. And so this result is statistically highly significant. Also, we surveyed “KEGG pathways in cancer” (http://www.genome.jp/kegg-bin/show_pathway?map=hsa05200&show_description=show) and found two genes including of *ZFP36L1* and *NF1* are novel driver genes in this pathways and they can be candidate and introduced as novel genomic biomarkers after more investigations and experiments. In the future we will apply other multiple bioinformatics tools as base classifiers and then we intend to develop an ensemble classifier machine learning method for metastatic breast cancers prognosis and diagnosis. However, to better evaluation of the accuracy of results, it is suggested that proper bio-molecular techniques such as qRT-PCR and NGS experiments are applied. Finally, this study can provide a road map to help researchers in using appropriate methods in the field of molecular biology and diagnosis of metastatic breast cancer.

Keywords: Somatic mutations; Breast invasive carcinoma; Driver genes; Bioinformatics tools; Prediction

References

[1] M. Ebrahimi, M. Vahdaninia, and A. Montazeri, “Risk factors for breast cancer in Iran: a case-control study.,” *Breast Cancer Res.*, vol. 4, p. 4, 2015.