Classification of large DNA methylation datasets for identifying cancer drivers

Fabrizio Celli, Fabio Cumbo, Emanuel Weitschek

PII: S2214-5796(17)30270-8
DOI: https://doi.org/10.1016/j.bdr.2018.02.005
Reference: BDR 86

To appear in: *Big Data Research*

Received date: 28 September 2017
Revised date: 12 February 2018
Accepted date: 28 February 2018

Please cite this article in press as: F. Celli et al., Classification of large DNA methylation datasets for identifying cancer drivers, *Big Data Res.* (2018), https://doi.org/10.1016/j.bdr.2018.02.005

ELSEVIER

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Classification of large DNA methylation datasets for identifying cancer drivers

Fabrizio Celli[a], Fabio Cumbo[a,b], and Emanuel Weitschek[c,a]

[a]*Institute of Systems Analysis and Computer Science, National Research Council, Via dei Taurini 19, 00185 Rome, Italy*

[b]*Department of Engineering, Roma Tre University, Via della Vasca Navale 79, 00154 Rome, Italy*

[c]*Department of Engineering, Uninettuno International University. Corso Vittorio Emanuele II, 39 00186 Rome, Italy*

Abstract

DNA methylation is a well-studied genetic modification crucial to regulate the functioning of the genome. Its alterations play an important role in tumorigenesis and tumor-suppression. Thus, studying DNA methylation data may help biomarker discovery in cancer. Since public data on DNA methylation become abundant – and considering the high number of methylated sites (features) present in the genome – it is important to have a method for efficiently processing such large datasets. Relying on big data technologies, we propose BIGBIOCL an algorithm that can apply supervised classification methods to datasets with hundreds of thousands of features. It is designed for the extraction of alternative and equivalent classification models through iterative deletion of selected features.

We run experiments on DNA methylation datasets extracted from The Cancer Genome Atlas, focusing on three tumor types: breast, kidney, and thyroid carcinomas. We perform classifications extracting several methylated sites and their associated genes with accurate performance (accuracy>97%). Results suggest that BIGBIOCL can perform hundreds of classification iterations on hundreds of thousands of features in few hours. Moreover, we compare the performance of our method with other state-of-the-art classifiers and with a wide-spread DNA methylation analysis method based on network analysis. Finally, we are able to efficiently compute multiple alternative classification models and extract - from DNA-methylation large datasets - a set of candidate genes to be further investigated to determine their active role in cancer. BIGBIOCL, results of experiments, and a guide to carry on new experiments are freely available on GitHub at https://github.com/fcproj/BIGBIOCL.

Keywords: classification; machine learning; DNA methylation; cancer; disease diagnostic predictive models; algorithms and techniques to speed up the analysis of big medical data.