proceedings

Proceedings

Gene Signatures Research Involved in Cancer Using Machine Learning †

Jose Liñares-Blanco and Carlos Fernandez-Lozano *

Department of Computer Science, Faculty of Computer Science, University of A Coruña, CITIC, A Coruña 15071, Spain; j.linares@udc.es

* Correspondence: carlos.fernandez@udc.es; Tel.: +34-881-01-6013

† Presented at the 2nd XoveTIC Conference, A Coruña, Spain, 5–6 September 2019.

Published: 31 July 2019

Abstract: With the cheapening of mass sequencing techniques and the rise of computer technologies, capable of analyzing a huge amount of data, it is necessary nowadays that both branches mutually benefit. Transcriptomics, in this case, is a branch of biology focused on the study of mRNA molecules, among others. The quantification of these molecules gives us information about the expression that a gene is having at a given moment. Having information on the expression of the approximately 20,000 genes harbored by human beings is a really useful source of information for the study of certain conditions and/or pathologies. In this work, patient expression -omic data data have been used to offer a new analysis methodology through Machine Learning. The results of this methodology were compared with a conventional methodology to observe how they differed and how they resembled each other. These techniques, therefore, offer a new mechanism for the search of genetic signatures involved, in this case, with cancer.

Keywords: machine learning; cancer; transcriptomics; TCGA; RNA-seq

1. Introduction

Having access to the expression of the whole spectrum of genes of an individual gives us the possibility to identify specific expression patterns of a condition and/or pathology. Nowadays, with the use of Machine Learning (ML) techniques, and with the large amount of data available for free, it is possible to use these techniques to extract new knowledge from the analysis. ML is able to identify expression patterns and/or gene subgroups that conventional techniques are not able to detect. For this reason, the detection of differential genetic expression patterns has been proposed for two groups of patients: patients with colon cancer and patients with lung cancer. The analysis will be carried out using two different approaches: conventional statistics and ML. Our work was published before in "Machine Learning Paradigms. Learning and Analytics in Intelligent Systems" [1].

2. Results

In this paper, a new way of analyzing gene expression data is proposed. The use of ML offer the possibility of widening the search space in terms of genes of interest. Unlike conventional analysis techniques, ML techniques allow working with hundreds and thousands of variables. The final objective of the analysis is to find those genes that have a differential expression between two patient populations, labeled as COAD and LUAD.

2.1. Conventional Analysis of Differential Gene Expression

In order to decide whether, for a given gene, there is a significant statistical difference in the number of mapped readings of that gene for different biological conditions, a statistical test should be performed, where the count of readings should be modeled to a certain distribution. Once the distribution that follow the data has been defined, and in this case, the dispersion of the data has been calculated, the differential expression of the transcripts is determined by the corresponding statistical tests for hypothesis contrast. Today there are different implementations in several statistical software that execute all this analysis in a simple way for the researcher. In this work we have used one of the most used in the field, called edgeR [2] package. According with a classical approach, Table 1 shows the 10 genes that have obtained the most significant values through classical analysis.

Table 1. Classical approach. ten genes with the higher impact between conditions.

Gene Name	*p*-Value
ITGA6	$2.605237 \times 10^{-245}$
AXIN2	$4.065388 \times 10^{-203}$
NOS2	$1.848360 \times 10^{-185}$
MYC	$4.409724 \times 10^{-171}$
TCF7	$3.930353 \times 10^{-163}$
COL4A3	$2.205117 \times 10^{-162}$
COL4A4	$1.548193 \times 10^{-138}$
TCF7L1	$2.527959 \times 10^{-110}$
PIK3R2	$6.857479 \times 10^{-103}$
BBC3	2.481885×10^{-99}

2.2. Data Analysis Using Machine Learning

On the other hand, the development of Machine Learning algorithms has greatly benefited the analysis of complex data, such as genomic data. In this work we have used ML to solve a classification problem (COAD vs LUAD), providing in this way, a new way to model transcriptomic data and thus to be able to extract new knowledge and search for new genetic signatures involved in cancer. The analysis of the importance of the variables gives us a fairly realistic approximation of what is happening. In Figure 1 we can see how the genes COL4A3 and NOS2 are the most important. On the other hand, PIK3CD and COL4A4 have hardly any weight in the model. If we compare the Top 10 genes of both approximations we observe coincidences in 7 genes and differences in 3 of them. As far as the conventional approximation is concerned, this presents significant results for the TCF7L1, PIK3R2 and BBC3 genes that the ML has not detected among the Top 10. For its part, ML techniques added NFKBIA, RASSF5 and PIK3CD genes among its Top 10.

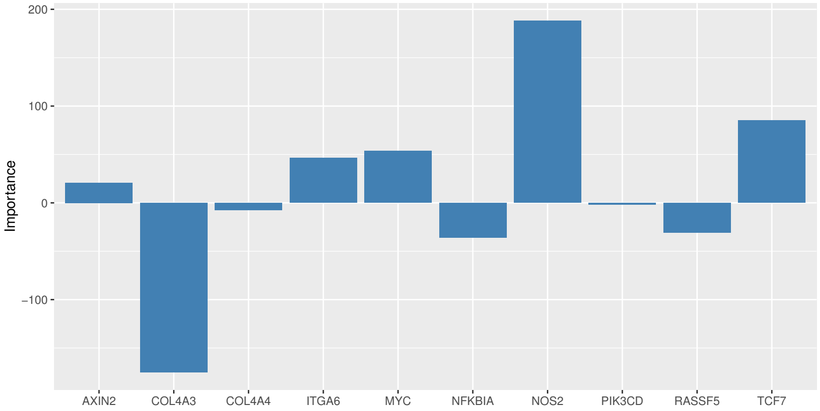

Figure 1. Variable importance according with the glmnet algorithm.

3. Discussion

The results obtained in this work indicate that ML offer coherent results in comparison with conventional techniques. The results that are observed shown that for a simple classification problem, both approaches reach almost the same results, although it is true that ML techniques may offer different possibilities when searching for new genetic marks. It is for this reason that the use of these techniques is considered useful when problems increase in complexity and the spectrum of genes involved in the pathology, such as cancer, is unknown.

4. Materials and Methods

The data has been downloaded from The Cancer Genome Atlas (TCGA) repository [3] from colon cancer patients (COAD) and lung cancer patients (LUAD). Due to the great dimensionality of the data (around 20,000), those genes belonging to specific cellular pathways were selected. In this case, genes were selected that had been previously identified in the routes related to colon cancer and lung cancer. For this purpose, the repository KEGG [4] was used, through the package KEGGREST [5] of R [6]. Specifically, the identifiers of pathways hsa05222, hsa05223 and hsa05210 were used, thus reducing the dimensionality to 173 genes. An univariate method (Kruskal test) were used to rank the genes. As for the classical approach, the edgeR package [2] has been taken as a reference. A Nested Cross Validation was used for training the models. In other words, there were two validation phases. Firstly, a holdout was used for the selection of the best hyperparameters (2/3 for training and 1/3 for testing) and secondly, a 10-fold CV was used for the validation of the model (we ran 5 times this CV process).

Author Contributions: Conceptualization, C.F.-L.; methodology, C.F.-L.; software, J.L.B. and C.F.-L.; formal analysis, J.L.B.; Writing—Original Draft preparation, J.L.B.; Writing—Review and Editing, J.L.B. and C.F.-L.; supervision, C.F.-L.

Funding: This research received no external funding.

Acknowledgments: This work is supported by the "Collaborative Project in Genomic Data Integration (CICLOGEN)" PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER)—"A way to build Europe". This project was also supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. ED431G/01, ED431D 2017/16), the "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23), and the Spanish Ministry of Economy and Competitiveness via funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union and the "Juan de la Cierva" fellowship program supported by the Spanish Ministry of Economy and Competitiveness (Carlos Fernandez-Lozano, Ref. FJCI- 2015-26071).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liñares Blanco, J.; Gestal, M.; Dorado, J.; Fernandez-Lozano, C. Differential Gene Expression Analysis of RNA-seq Data Using Machine Learning for Cancer Research. In *Machine Learning Paradigms. Learning and Analytics in Intelligent Systems*; Springer: Cham, Switzerland, 2019; pp. 27–65.
2. McCarthy, D.J.; Chen, Y.; Smyth, G.K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **2012**, *40*, 4288–4297.
3. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **2015**, *19*, A68.
4. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30.
5. Tenenbaum, D. KEGGREST: Client-side REST access to KEGG. *R Package Vers.* **2016**, *1*, DOI: 10.18129/B9.bioc.KEGGREST.
6. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.