proceedings

MDPI

Proceedings

Predicting Gastric Cancer Molecular Subtypes from Gene Expression Data [†]

Marta Moreno [1,2,*], Abel Sousa [3,4,5,6], Marta Melé [7], Rui Oliveira [2,8] and Pedro G Ferreira [1,2,3,4,*]

[1] Department of Computer Science, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
[2] University of Minho and INESC TEC, 4200-465 Porto, Portugal; rui.oliveira@inesctec.pt
[3] Ipatimup—Institute of Molecular Pathology and Immunology of the University of Porto, 4200-465 Porto, Portugal; abels@ipatimup.pt
[4] i3s—Instituto de Investigação e Inovação em Saúde da Universidade do Porto, 4200-135 Porto, Portugal
[5] Graduate Program in Areas of Basic and Applied Biology, Abel Salazar Biomedical Sciences Institute, University of Porto, 4050-313 Porto, Portugal
[6] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK
[7] Life Sciences Department, Barcelona Supercomputing Center, Barcelona, 08034 Catalonia, Spain; marta.mele@bsc.es
[8] Department of Informatics, University of Minho, 4710-057 Braga, Portugal
[*] Correspondence: up201305416@fc.up.pt (M.M.); pgferreira@fc.up.pt (P.G.F.)

[†] Presented at the 3rd XoveTIC Conference, A Coruña, Spain, 8–9 October 2020.

Published: 7 September 2020

Abstract: Stomach cancer is a complex disease and one of the leading causes of cancer mortality in the world. With the view to improve patient diagnosis and prognosis, it has been stratified into four molecular subtypes. In this work, we compare the results of multiple machine learning algorithms for the prediction of stomach cancer molecular subtypes from gene expression data. Moreover, we show the importance of decorrelating clinical and technical covariates.

Keywords: gene expression; gastric cancer; disease classification; machine learning

1. Introduction

Several large-scale projects, such as TCGA (The Cancer Genome Atlas) or ICGC (International Cancer Genome Consortium), have studied dozens of tumor types through the analysis of hundreds of samples with several molecular assays of the genome, epigenome, proteome, transcriptome and the respective clinical data. One such example is stomach adenocarcinoma (STAD), representing nearly 5% of new cancer cases worldwide [1]. STAD is a complex disease, with a mortality rate almost equivalent to its incidence.

The molecular profiling of more than four hundred tumor cells with five different assays has allowed for the identification of four novel STAD sub-types with different diagnostic and prognostic value [2]. However, extensive characterization of tumor samples is not always possible due to clinical, technical or budget limitations.

Previous studies have shown that strong outcome predictor signatures can be derived from RNA data in cancer [3]. These studies indicate that gene expression carries sufficient signal for the accurate prediction of phenotypes. For this reason, we believe that the genetic alterations observed in different STAD molecular subtypes should be reflected in differential tissue gene expression

Here, we set to investigate if it is to possible to develop a predictive tool that, based on transcriptome profiling with RNA-seq, can predict stomach cancer samples according to the proposed stratification. In order to minimize the effect of possible unwanted sources of variation in the data, we have analyzed the impact of pre-processing the data, taking into account the effect of the available covariate information.

2. Materials and Methods

STAD-specific transcriptome data were obtained from the TCGA Research Network (https://www.cancer.gov/tcga). Samples with insufficient clinical information were excluded. As features, only coding genes with a median Fragments per Kilobase per Million (FPKM) value higher than 1 were retained (Figure 1) and their values were log2 transformed.

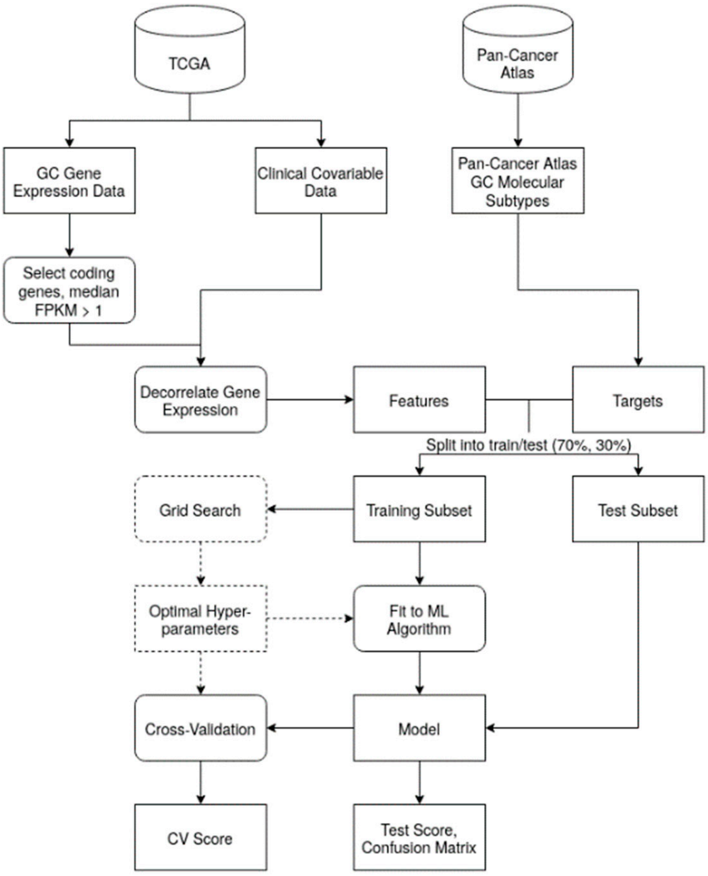

Figure 1. Diagram of the pipelines used in this work. Steps with a dashed line were only performed on pipelines with hyper-parameter optimization.

Technical or clinical factors may correlate with both the features and the target STAD molecular subtypes, possibly confounding machine learning (ML) predictions. Without a decorrelation step, the model may thus over- or under-estimate the effect of the features on the target variable. As a data pre-processing step, we regressed out the possible confounding effects of the covariates on the gene expression data through a multiple linear model:

$$g_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{age} + \boldsymbol{\beta}_2 \text{gender} + \boldsymbol{\beta}_3 \text{race} + \boldsymbol{\beta}_4 \text{age\_diagnosis} + \boldsymbol{\beta}_5 \text{distant\_metastasis} + \boldsymbol{\beta}_6 \text{primary\_tumor} + \boldsymbol{\beta}_7 \text{icd-10} + \boldsymbol{\beta}_8 \text{morphology} + \boldsymbol{\beta}_9 \text{diagnosis} + \boldsymbol{\beta}_{10} \text{prior\_malignancy} + \boldsymbol{\beta}_{11} \text{tissue} + \boldsymbol{\beta}_{12} \text{tumor\_stage} + \circledcirc$$

where g_i represents the gene expression for gene i, $\boldsymbol{\beta}_0$ is the intercept, $\boldsymbol{\beta}_i$ $i \in (1, \ldots, 12)$ is the regression coefficients for the covariates, and $\circledcirc$ is the noise term.

The residuals of the model, obtained as the difference between the real gene expression value (gi) and the predicted expression ($\hat{g}$i), were used as the expression phenotype.

After this step, several ML pipelines were devised with the goal of predicting STAD molecular subtypes from RNA-seq data (chromosomal instability (CIN) 61.45%, Epstein–Barr virus (EBV) 7.54%, genomically stable (GS) 12.85%, microsatellite instability (MSI) 18.16%; see Figure 2a). First, the dataset was split into stratified training (n = 250) and test (n = 108) sets. Each algorithm learned from the training set's features to build prediction models, with or without hyper-parameter optimization. Cross-validation was performed to test the model's performance on sampled portions of the training data, with subsequent validation using the unseen test set.

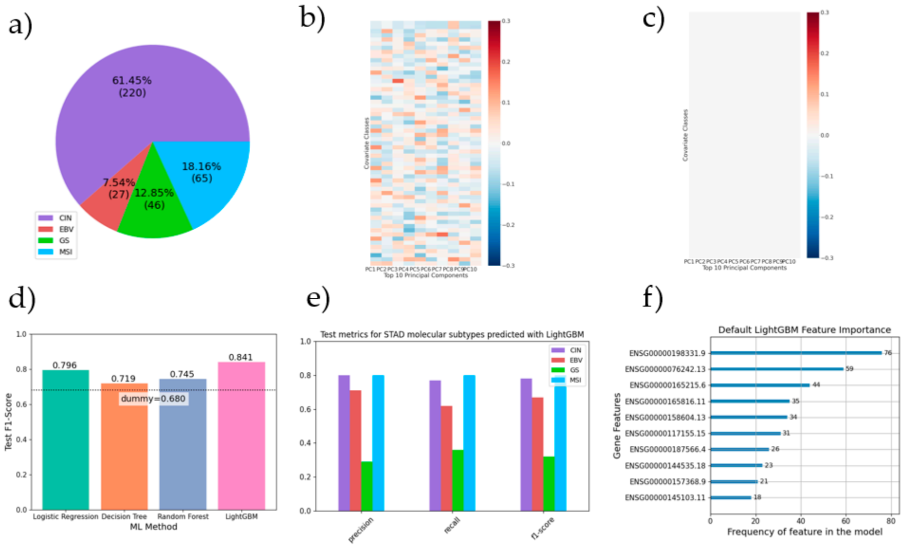

Figure 2. (**a**) Distribution of STAD molecular subtypes in the data (N = 358). (**b**,**c**) Correlation heatmap between clinical covariates and the top 10 gene expression principal components. (**b**) Before covariate decorrelation. (**c**) After covariate decorrelation. (**d**) Distribution of test f1-scores across methods, as compared to the dummy estimator's (which always predict most frequent class) score. (**e**) Test metrics obtained using LightGBM with default settings, stratified by class. (**f**) The top 10 gene features by their importance for the LightGBM default model.

3. Results

Several covariates possessed significant correlation (ranging from -0.14 to 0.17) with the top 10 principal components for gene expression (Figure 2b). As expected, all covariate correlation was lost after gene-wide covariate decorrelation (Figure 2c).

Despite a heavy class imbalance (Figure 2a), all machine learning models outperformed a dummy estimator that always predicted the most frequent class, with an average 8% improvement across methods (Figure 2d). There were also notable differences in performance between algorithms, with the best performer, LightGBM, having a test F1-score 5.6% better than the second best, logistic regression. By contrast, there was no significant difference between results of models using default algorithm hyper-parameters and those obtained following hyper-parameter optimization. On a per class basis, the CIN sub-type exhibits the best results (Figure 2e). The top 10 most informative gene features for the best performing model (LightGBM default) are shown in Figure 2f. Of special interest, the second most contributing gene, ENSG00000076242 (MLH1), is a tumor suppressor gene whose epigenetic silencing is associated to MSI tumors.

4. Discussion

Machine learning methods show promise for the prediction of molecular subtypes in STAD, with even the simplest methods performing better than random chance. However, perhaps due to the small sample size and/or imbalance of the data, hyper-parameter optimization offered no performance improvements.

Funding: This work was supported by the FCT (Fundação para a Ciência e a Tecnologia) research grant Ph.D. Studentship SFRH/BD/145707/2019 and the research grant IF/01127/2014, funded in the scope of the FCT Investigator Exploratory Project: "Understanding the impact of acquired and germline genetic variants in the complexity of gastric cancer"; GenomePT project (reference 22184): "National Laboratory for Genome Sequencing and Analysis"; QREN L3 project (reference NORTE-01-0145-FEDER-000029): "Mapping genetic and phenotypic heterogeneity in HER2 positive cancers to anticipate and counteract resistance phenotypes".

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

References

1. World Health Organization; International Agency for Research on Cancer (IARC). *GLOBOCAN 2018: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2018*; WHO: Geneva, Switzerland; IARC: Lyon, France, 2018. Available online: https://gco.iarc.fr/today/online-analysis-pie (accessed on 23 July 2020).
2. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **2014**, *513*, 202–209, doi:10.1038/nature13480.
3. Byron, S.A.; Van Keuren-Jensen, K.R.; Engelthaler, D.M.; Carpten, J.D.; Craig, D.W. Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nat. Rev. Genet*. **2016**, *17*, 257–271, doi:10.1038/nrg.2016.10.