DSSV 2023 Data Science, Statistics & Visualisation
ECDA 2023 European Conference on Data Analysis

Evaluating the impact of inclusion of molecular information in glioma classification on network-based biomarker discovery

R. Coletti[1, 2], M. L. Mendonça[3], S. Vinga[3, 4], and M. B. Lopes[1,2,5]

[1]NOVA School of Science and Technology (FCT NOVA), Universidade NOVA de Lisboa, Caparica, 2825-149 Portugal.
[2]Center for Mathematics and Applications (NOVA MATH), FCT NOVA, Caparica, 2825-149 Portugal.
[3]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, R. Alves Redol 9, Lisbon, 1000-029, Portugal.
[4]IDMEC, Instituto Superior Técnico, Universidade de Lisboa, R. Av. Rovisco Pais 1, Lisbon, 1049-001, Portugal.
[5]Research and Development Unit of Mechanical and Industrial Engineering (UNIDEMI), FCT NOVA, Caparica, 2825-149 Portugal.

Glioma is a heterogeneous group of brain tumor presenting three main types that can exhibit different prognoses. Glioma-type identification depends on specific parameters defined by the official World Health Organization (WHO) classification of the Central Nervous System (CNS), which is constantly updated to support the diagnostic process. Indeed, while for years glioma diagnoses have been exclusively based on histology, recently the introduction of genetic and molecular criteria significantly changed the classification procedures [1].

In this context, searching for new potential diagnostic and prognostic biomarkers is crucial, and mathematical and statistical tools can be effectively employed to infer important information from the large datasets nowadays available. Based on a RNA-Sequencing dataset from The Cancer Genome Atlas (TCGA) data portal, with glioma diagnosis updated according to the 2016 and 2021 WHO guidelines ($\sim$ 20K variables), we propose a pipeline for biomarker discovery, aiming at identifying glioma-type specific genes with diagnostic and prognostic value (Figure 1). Our methodology first applies the graphical lasso method [2] to perform a network-based variable selection for each glioma type. The selected genes are analyzed in order to identify key features in the estimated networks with potential diagnostic value. The identified subsets of variables are then validated and further investigated through survival analysis, performed with regularized Cox regression modeling [3], to determine which of these selected genes also carry prognostic information. All these results are integrated into a network representation, which we propose as a functional tool to disclose unknown relations and support biological research.

Overall, our study identifies potential biomarkers characteristic of each glioma type, yet leading to better results considering the 2021 WHO classification, which is mainly built upon molecular information. However, our survival analysis point out that histology also has an impact on the predicted risk of death, suggesting that additional efforts are needed to further characterize the heterogeneity of glioma types and support the classification procedure.

[1] T. R. Berger, P. Y. Wen, M. Lang-Orsini, U. N. Chukwueke. "World Health Organization 2021 Classification of Central Nervous System Tumors and Implications for Therapy for Adult-Type Gliomas: A Review", *JAMA oncology*, vol. 10, pp. 1493–1501, 2022.

[2] J. Friedman, T. Hastie, R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso ", *Biostatistics*, vol. 9, pp. 432–441, 2008.

[3] D.R. Cox, "Regression Models and Life-Tables", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, pp. 187-202.
R. Tibshirani. "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society. Series B (Methodological)* vol. 58, pp. 267–88, 1996.

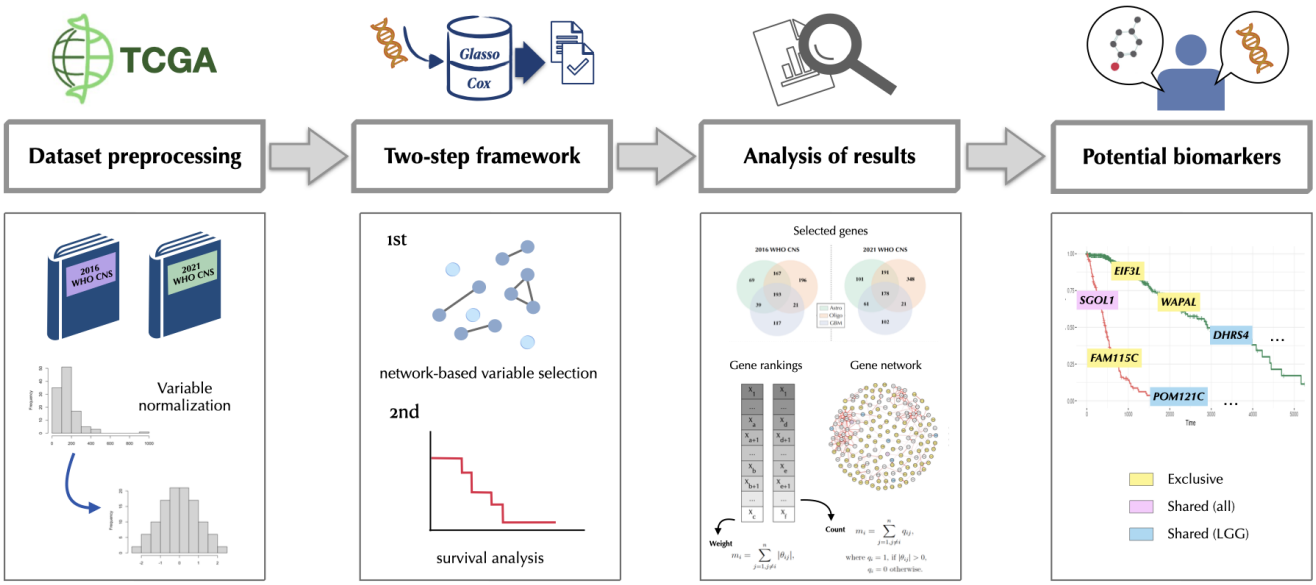

Figure 1: Workflow of our analysis. Dataset preprocessing: the TCGA RNASeq dataset has been updated according to the 2016 and 2021 WHO classifications, and the variables have been normalized to apply the graphical lasso algorithm. Two-step framework: after a network-based variable selection by graphical lasso, the results have been validated through regularized Cox regression survival analysis. Analysis of results: for each step, the results have been analyzed in light of the two 2016 and 2021 WHO classifications. Potential biomarkers: the outcomes of our methodology led to a list of potential biomarkers that could be exclusively selected for a given glioma type, or shared.