Term Project – Phase 1: Proposal

Team Members: Gavrielatos Marios | 7115152100023
Dimitra Paranou | 7115152100034

Paper: Cheng, Quan, Jing Li, Fan Fan, Hui Cao, Zi-Yu Dai, Ze-Yu Wang, and Song-Shan Feng. 2020. "Identification and Analysis of Glioblastoma Biomarkers Based on Single Cell Sequencing." Frontiers in Bioengineering and Biotechnology 8 (March): 167, https://doi.org/10.3389/fbioe.2020.00167.

Data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84465

1. What is the problem studied in the paper you selected?

Glioblastoma (GBM) is an aggressive type of cancer that can occur in the brain or spinal cord. The survival time after diagnosis is 12-15 months, with 5-year survival rate <5% (Stupp et al. 2017, 2005). Studies have shown that there are three molecular mechanisms which lead to GBM development, mutations activate the receptor of the tyrosine kinase (RTK) gene which lead to dysregulation of growth factor signaling, activation of the phosphatidyl inositol 3-kinase (PI3K) signaling and finally deactivation of the p53 tumor protein and retinoblastoma tumor suppressor pathways (Furnari et al. 2007). Furthermore, gene expression experiments identified four distinct subtypes of GBM, Neural, Proneural, Classical and Mesenchymal (Verhaak et al. 2010).

Cancer is a dynamic disease which progresses during its development and becomes more heterogeneous. Many tumors of the same type are characterized by heterogeneity of their genes and proteins. Glioblastoma heterogenesis makes cancer therapies ineffective as heterogenous GBM cells are combined with the complex tumor microenvironment (Becker et al. 2021). For this reason, it is highly important for researches to be able to profile the different cell types which are located in the tumor ecosystem in order to develop specialized treatments. Single-cell transcriptomic analysis of the tumor enables researchers to assay a great number of individual cells.

2. Why it is important? What is the state of the art in this topic?

Biomarkers are cellular or molecular alterations that are "cellular, biochemical or molecular alterations that are measurable in biological media such as human tissues, cells, or fluids" as described by (Hulka and Wilcosky 1988). Biomarkers can provide information about the dynamic of a disease which may lead to the understanding of various diseases, to faster diagnosis and to the development of a treatment in diseases like GBM (Mayeux 2004).

Single cell RNA sequencing analysis helped identify genetic molecular mechanisms in patients with GBM by analyzing the gene expression of the tumor and its microenvironments. Several studies have also identified potential biomarkers to distinguish the different lineages of the immune cells that are located GBM. In this study, the authors analyzed scRNAseq data in order to find core biomarkers that are able to distinguish the discrepancy between GBM tumor cells and cells from the pericarcinomatous environments (Cheng et al. 2020).

3. What are the main data analysis methods and tools used, and the main results claimed by the authors?

The authors used data from 2343 tumors cells and 1246 peripheral cells in order to perform maximum relevance minimum redundancy (mRMR) analysis. mRMR analysis is used to identify characteristics of genes and phenotypes. Maximum relevance is a feature selection technique which aims to characterize redundant features which must be removed. By performing mRMR the authors selected the top 100 genes. Moreover, they created 100 SVM classifiers and performed incremental feature selection (IFS) in order to identify the minimum/optimal number of genes. The leave-one-out cross validation (LOOCV) was used to evaluate the prediction performance of each SVM. Finally, Mathew's correlation coefficient (MCC) was used in IFC optimization. The peak MCC corresponded to 31 genes which were selected as the minimum number of GBM biomarkers.

In the next step, the authors performed t-distribution stochastic neighbor embedding (t-SNE) in order to test the accuracy of the classification since tumors usually consists of tumor cells and normal cells. The plots showed that GBM tissues may contain non-GBM cells and non-GBM tissues may contain GBM cells.

The authors performed Gene Ontology (GO) enrichment analysis for the 31 genes in order to find which genes may serve as potential therapeutic targets.

4. Motivation for additional analysis

Our scope is to further analyze data to reveal potential prognostic GBM marker genes, elaborate their functions, and build a prognostic model for GBM patients.

The first goal of this project is to reproduce the paper's methods and verify the authors claims. Cheng et. al, start by identifying the differentially expressed genes (DEGs) and then construct Support Vector Machines to classify the DEGs, applying an incremental feature selection (IFS) method. Then, the team uses t-distributed stochastic neighbor embedding (t-SNE) plots of predicted GBM cells and predicted non-GBM cells to check misclassifications.

We are going to perform an analysis in order to find the best model possible for DEG classification. This analysis is going to be comprised of 2 steps. During the first step, we aim to select the best model family (Random Forests, XGBoost, SVM, Naive Bayes) by using nested cross validation. We will perform hyperparameter optimization for every model family in order to find the best model. This process will give us the average test performance of every model family. Moreover, we are going to store the hyperparameters from every model instance. On the second step of the pipeline, we will perform hyperparameter tuning on the best model instance from step one. The training is done using the whole dataset and we will use cross-validation. Finally, we will test the effect of ensemble modeling in order to try to achieve better predictive performance. This analysis will be performed in order to improve the current results and test the effectiveness of different classification methods.

After completing the above procedure and depending on time, we want to approach the data in a different way and focus on the prognostic values of GBM data. The workflow that we plan to follow is described below:

1. To begin with, we will start the workflow by identifying the differentially expressed genes (DEGs) between GBM and normal brain tissues. Using clustering after normalization of data will eliminate abnormal samples and keep only the desired samples.

2. Then, using p-value and logFC values, we will scrutinize our values and further assess the accuracy of DEGs. Principal component analysis will also be applied to categorize the data.
3. Next step is the identification of prognostic values of DEGs and implementation of survival analysis. For this task, univariate Cox proportional hazard regression and LASSO regression will be used. The former is essentially a regression model commonly used statistically in medical research for investigating the association between the survival time of patients and one or more predictor variables, while the latter is a type of linear regression that uses shrinkage and is a useful method to determine interpretable prediction rules in high dimension data (Y. Li et al. 2019).
4. To determine their prognostic value, Kaplan-Meier plots will be used with $p < 0.05$ (statistically significant). The Kaplan-Meier estimate is also called “product limit estimate”. It involves computing probabilities of occurrence of events at a certain point of time.
5. Finally, a time-dependent receiver operating characteristic (ROC) curve and area under the curve (AUC) will be implemented to evaluate the prediction accuracy of the risk model and the selected genes.

The extending approach is a common prognostic analysis of data for gene/cell identification and estimation of the survival rates (H. Li and Gui 2004). As significant morphological, clinical and biological prognostic factors vary according to molecular subtypes of tumors, yet comprehensive analysis of such factors linked to survival in each group is lacking. So, it will be a really fruitful research for us, as it will introduce us to this kind of research, and a good opportunity to analyze and examine the survival rate of patients with GBM.

5. Technical approach

Concerning the technical approach, both of the programming languages - Python and R - will be used. For the reproduction of paper, we will use Python, while for the survival analysis R is more powerful. The main libraries that we are going to import are describing below:

Python

1. Mrmr / pymrmr (for feature selection)
2. Sklearn (SVM, t-SNE, PCA, Kmeans, LASSO, tree, RandomForestClassifier)
3. Lifelines (KaplanMeierFitter, COX)
4. XGboost

R

1. Survival
2. Enrichplot
3. survivalROC
4. DealGPL570

The Machine Learning methods that we are going to use are described below. In the first task, where we will reproduce the results, Support Vector Machine will be used to classify the data. SVMs are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

For our additional analysis, Random Forest Classifier, XGBoost, Decision trees and Naive Bayes will be implemented. Then for the survival analysis, we will use LASSO, which is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model, and COX regression to evaluate

simultaneously the effect of several factors on survival. Then, other ML methods will take place, like dimensionality reduction techniques (PCA, t-SNE) for transforming to low dimensions of data retaining some meaningful properties of the original data and Clustering (Kmeans).

6. Implementation plan

The planning of the project is separated into two parts. The first part includes the reproduction of the project, where the main goal is to implement the methods that the authors follow and verify their claims. The second part consists of the additional analysis, where families of classification models will be evaluated in order to select the optimum one and finally survival analysis will be implemented. The steps and the team member roles are presented below:

- The reproduction of the selected analysis will be performed by both members.
    1. Data Fetching - both members
    2. Identification of differentially expressed genes / Feature selection (mRMR)
    3. Classification (Support Vector Machine)
    4. Check of misclassification (t-SNE)

- Addition work
    1. Nested Cross Validation in different family models. Each member will evaluate two model families.
    2. Select the one with the best performance.
    3. Using the whole dataset, find the optimal parameters for optimizing the model.
    4. Perform ensemble modeling.
    5. Identification of differentially expressed genes / Normalization / Clustering / Filtering
    6. Investigate the relationship between overall survival (OS) and gene expression levels (COX & LASSO Regression).
    7. Determine the prognostic value (Kaplan-Meier).
    8. Evaluate the prediction accuracy of the risk model and the selected genes (ROC - AUC).
    9. Evaluate and comment on the results.

7. References

Becker, Aline P., Blake E. Sells, S. Jaharul Haque, and Arnab Chakravarti. 2021. “Tumor Heterogeneity in Glioblastomas: From Light Microscopy to Molecular Pathology.” *Cancers* 13 (4). https://doi.org/10.3390/cancers13040761.

Cheng, Quan, Jing Li, Fan Fan, Hui Cao, Zi-Yu Dai, Ze-Yu Wang, and Song-Shan Feng. 2020. “Identification and Analysis of Glioblastoma Biomarkers Based on Single Cell Sequencing.” *Frontiers in Bioengineering and Biotechnology* 8 (March): 167.

Furnari, Frank B., Tim Fenton, Robert M. Bachoo, Akitake Mukasa, Jayne M. Stommel, Alexander Stegh, William C. Hahn, et al. 2007. “Malignant Astrocytic Glioma: Genetics, Biology, and Paths to Treatment.” *Genes & Development* 21 (21): 2683–2710.

Hulka, B. S., and T. Wilcosky. 1988. “Biological Markers in Epidemiologic Research.” *Archives of Environmental Health* 43 (2): 83–89.

Li, Hongzhe, and Jiang Gui. 2004. “Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data.” *Bioinformatics* 20 Suppl 1 (August): i208-15.

Li, Yin, Di Ge, Jie Gu, Fengkai Xu, Qiaoliang Zhu, and Chunlai Lu. 2019. “A Large Cohort Study Identifying a Novel Prognosis Prediction Model for Lung Adenocarcinoma through Machine Learning Strategies.” *BMC Cancer* 19 (1): 886.

Mayeux, Richard. 2004. “Biomarkers: Potential Uses and Limitations.” *NeuroRx: The Journal of the American Society for Experimental NeuroTherapeutics* 1 (2): 182–88.

Stupp, Roger, Warren P. Mason, Martin J. van den Bent, Michael Weller, Barbara Fisher, Martin J. B. Taphoorn, Karl Belanger, et al. 2005. “Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma.” *The New England Journal of Medicine* 352 (10): 987–96.

Stupp, Roger, Sophie Taillibert, Andrew Kanner, William Read, David Steinberg, Benoit Lhermitte, Steven Toms, et al. 2017. “Effect of Tumor-Treating Fields Plus Maintenance Temozolomide vs Maintenance Temozolomide Alone on Survival in Patients With Glioblastoma: A Randomized Clinical Trial.” *JAMA: The Journal of the American Medical Association* 318 (23): 2306–16.

Verhaak, Roel G. W., Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, et al. 2010. “Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1.” *Cancer Cell* 17 (1): 98–110.