4th International Iranian Conference on BIOINFORMATICS
Faculty of Engineering, University of Zanjan
February 4-6 ,2025

ICB13
Iranian Conference on Bioinformatics
سیزدهمین همایش ملی بیوانفورماتیک ایران
دانشکده مهندسی دانشگاه زنجان ـ ۱۶ تا ۱۸ بهمن ۱۴۰۳

Identification of Driver Genes in Glioblastoma Based on Single-Cell Gene Expression Data Using Integrated Pseudotime and Phylogenetic Analysis

Fateme Mirza – Abolhassani [1], Sobhan Ahmadian Moghadam[2], Fatemeh Zare – Mirakabad[2], Kaveh Kavousi[3,*]

[1] Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran
[2] Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran
[3] Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

Presenter: Fateme.AbolHassani91@gmail.com

Introduction

Understanding the origins of cancer and the pathways driving its growth, particularly those involved in mitotic processes, remains a fundamental challenge in cancer research. Glioblastoma (GBM), the most aggressive primary brain tumor, exemplifies this challenge, exhibits notable intratumoral heterogeneity and genetic complexity, which make effective treatment highly obscure. Recent researches have approached this subject from two distinct perspectives: gene expression profiles and genetic mutations. While each has provided valuable insights, integrating these viewpoints may offer a more comprehensive understanding of GBM progression. In this study, we build a model that considers both types of data and identifies relevant results. Our findings demonstrating that mutations not only drive cancer progression but also provide deeper insights into the pathways underlying GBM evolution (Fig. 1.).

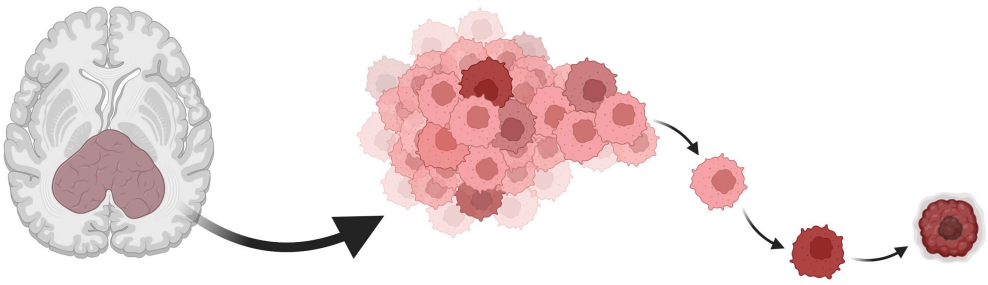

Figure 1: *Illustration of glioblastoma (GBM) progression, highlighting tumor growth, cellular evolution, and genetic complexity.*

Methods

We developed a model to investigate the relationship between gene expression and mutations in GBM, aiming to identify key genes involved in its development and progression. The input for this method consists of a set of cells and a set of initial genes related to GBM. The output is the identification of the genes that are most crucial in the development and progression of GBM. First of all, data was preprocessed (Fig. 2). Then, we built a phylogenetic tree using the neighbor-joining (NJ) algorithm with ClustalW as distance metric to analyze evolutionary relationships based on nucleotide sequences. A virtual root node was introduced to represent the direction of mutation propagation (Fig. 3). The Monocle algorithm was applied to order cells along pseudo-time paths based on their gene expression profiles. Cells were clustered into distinct developmental paths, representing progression through the biological process. We then compared the phylogenetic tree and pseudotime paths using various approaches, including entropy analysis, the construction of a semi-expression matrix, and the Weighted Mean Score (Fig. 3).

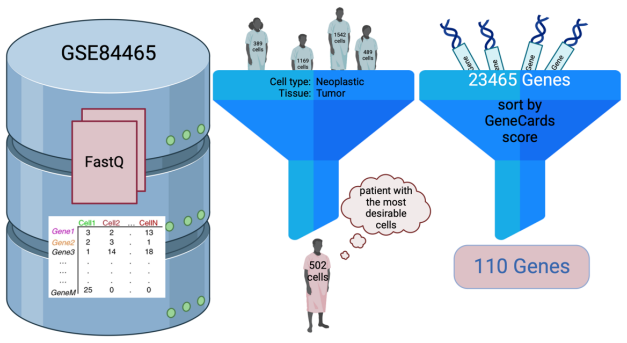

Figure 2: *Workflow for data preprocessing*

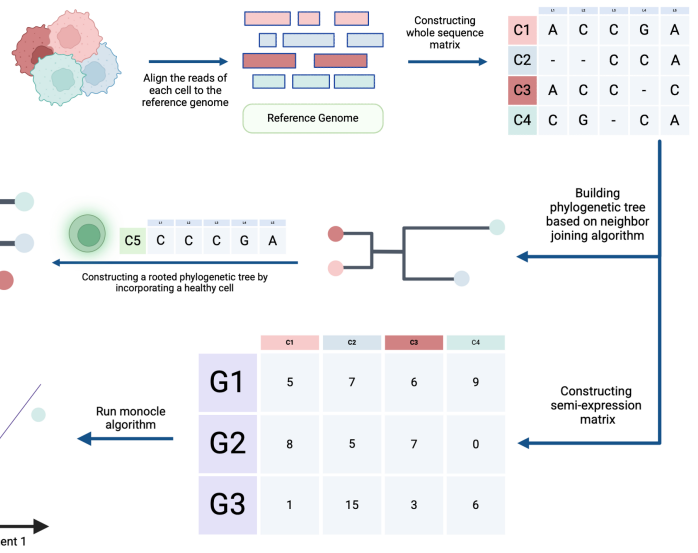

Figure 3: *Visualization of the workflow for constructing a rooted phylogenetic tree and a semi-expression matrix.*

Result

We constructed semi-expression and expression matrices using the top genes. Using the Monocle algorithm, we demonstrated that the semi-expression matrix captured distinctions among 502 cells with greater accuracy than the full expression matrix (Fig 4.). Among these, EGFR consistently emerged as a key gene across multiple paths. For instance, in the semi-expression matrix, cells in the initial positions of specific paths showed higher mutation percentages in genes like EGFR, COL4A6, and HIF1A, suggesting their significance in pathway differentiation. Additionally, we constructed a phylogenetic tree using these genes, colored by pseudotime paths derived from the expression matrix, resulting in a Weighted Shannon Entropy of 1.894 and a Weighted Mean Score of 29.639.

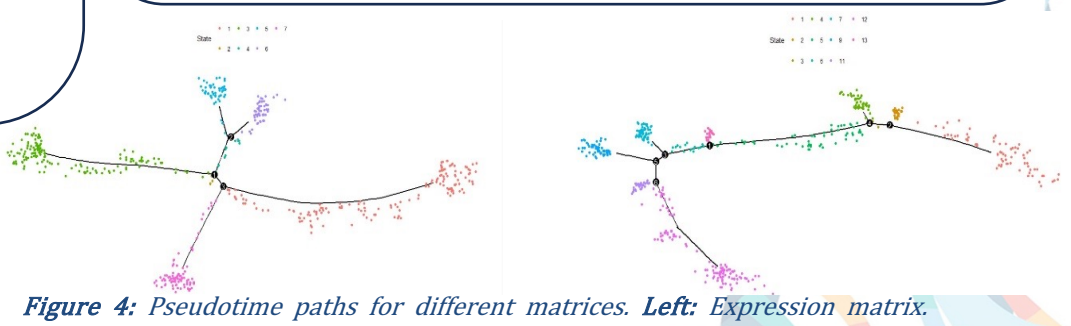

Figure 4: *Pseudotime paths for different matrices.* ***Left:*** *Expression matrix.* ***Right:*** *Semi-expression matrix. The right figure captures distinctions among 502 cells with greater accuracy, showing more detailed.*

References

[1] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S.,and Rinn, J.L. (2014) 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', Nature Biotechnology, 32(4), pp. 381–386. doi: 10.1038/nbt.2859.