Harini Kethar, Thomas Kidu, Elizabeth Tarangelo

Project Proposal: Predicting Colorectal Cancer Patient Outcomes from IO Therapy using Machine Learning based on MSI Signatures and Gene Mutation Status

Colorectal cancer is the third most common cancer to be diagnosed in the United States, affecting nearly 135,430 new patients and leading to 50,260 deaths as of 2017 [1]. The development and progression of CRC is driven by environmental factors, genetic mutations, and epigenetic mutations that continue to evolve and evade the body. While surgery, chemotherapy, and radiation therapy are the standard of care/first-line options presented, immunotherapy is a promising strategy that has emerged for many CRC patients [2].

Immunotherapies such as immune checkpoint blockades that target inhibitory pathways are able to take advantage of the body's immune system to recognize and attack tumor cells that are trying to avoid detection. The success rate of these therapies, evaluated by tumor progression and relapse rate, is correlated with expression of two genetic markers involved in early-stage metastatic tumor development: microsatellite instability (MSI) and KRAS proto-oncogene mutations [3]. MSI is a hypermutation that results from mismatch repair (MMR) that leads to an accumulation of mutations in the genome. Almost 15% of CRC cases exhibit MSI [4]. Mutational signatures within CRC reflect the biological processes that are contributing to tumor development and can be linked to specific subtypes of tumor classification as well as therapy response [5]. By integrating analysis of mutational signatures and MSI status with clinical outcomes, the tumor microenvironment can be better understood in order to understand how these genetic factors can be used in early cancer surveillance and be used to predict patient survival rates.

Our final project will focus on screening MSI and KRAS signature data [6] from patients diagnosed with stage I-IV adenocarcinomas who have been treated with checkpoint inhibitor therapy. Genomic and mutational sequencing data sets will be obtained from TCGA using SigProfiler to identify mutational signatures from whole-exome sequencing (WES) and whole-genome sequencing (WGS) data [7]. Patients will be stratified into cohorts based on expression of mutations and stage of disease progression using TCGA workflow and pymaf. Clinical outcomes of patient survival analysis after checkpoint inhibitor therapy will be performed using the Kaplan-Meier analysis and Cox regression analysis.

The project will be executed using Python and several bioinformatics tools to facilitate the extraction of mutational signatures from sequencing data [8]. Mutation files will be parsed and managed using pymaf and pyvcf. Pandas and NumPy will handle data preprocessing, normalization, and feature engineering. First, we will analyze and study the key features. These include analyzing the MSI-related metrics such as MSI Score, representing the fraction of unstable microsatellite loci, and subsequent MSI Status indicating binary classification (MSI-High or MSI-Low) [9]. Gene expression levels of checkpoint inhibitors Programmed Death-1 (PD-1) and cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) will be considered in this. Moreover, we will identify the key driver gene mutations, specifically binary indicators for mutations in significant cancer-associated genes (KRAS) and immune cell infiltration scores (such as T-cell and CD8+ cell infiltration) will be considered in relation to immunotherapy response. After selecting the key features as inputs for the predictive model, we are considering two approaches for designing the machine learning model: either solely utilizing tree-based models, such as Random Forest and XGBoost, or implementing a deep learning neural network model [10]. The final model will depend on the quality and size of the datasets available after the final preprocessing step. The goal is to predict the progression-free response on immunotherapy in correlation with the expression of MSI and KRAS mutational signatures.

References:

[1] A. Smith, B. Johnson, and C. Lee, *A Practical Guide to Biomarkers for the Evaluation of Colorectal Cancer*. New York, NY, USA: Springer, 2022. https://www.sciencedirect.com/science/article/pii/S0893395222009565

[2] H. Fadlallah, J. El Masri, H. Fakhereddine, J. Youssef, C. Chemaly, S. Doughan, and W. Abou-Kheir, "Colorectal cancer: Recent advances in management and treatment," *World Journal of Clinical Oncology*, vol. 15, no. 9, pp. 1136–1156, Sep. 2024, doi: 10.5306/wjco.v15.i9.1136. https://www.wjgnet.com/2218-4333/full/v15/i9/1136.htm

[3] G. M. Nash, M. Gimbel, A. M. Cohen, Z.-S. Zeng, M. I. Ndubuisi, D. R. Nathanson, J. Ott, F. Barany, and P. B. Paty, "KRAS Mutation and Microsatellite Instability: Two Genetic Markers of Early Tumor Development That Influence the Prognosis of Colorectal Cancer," *Annals of Surgical Oncology*, vol. 17, no. 2, pp. 416–424, Feb. 2010, doi: 10.1245/s10434-009-0713-0 https://pmc.ncbi.nlm.nih.gov/articles/PMC4380015/#abstract1

[4] G. Bogani, B. J. Monk, M. A. Powell, S. N. Westin, B. Slomovitz, K. N. Moore, R. N. Eskander, F. Raspagliesi, M.-P. Barretina-Ginesta, N. Colombo, and M. R. Mirza, "Adding immunotherapy to first-line treatment of advanced and metastatic endometrial cancer," *Annals of Oncology*, vol. 35, no. 5, pp. 414–428, May 2024, doi: 10.1016/j.annonc.2024.02.006. https://pubmed.ncbi.nlm.nih.gov/38431043/

[5] L. B. Alexandrov *et al.*, "The repertoire of mutational signatures in human cancer," *Nature*, vol. 578, pp. 94–101, 2020, doi: 10.1038/s41586-020-1943-3. https://www.nature.com/articles/s41586-020-1943-3

[6] G. M. Nash, M. Gimbel, A. M. Cohen, Z.-S. Zeng, M. I. Ndubuisi, D. R. Nathanson, J. Ott, F. Barany, and P. B. Paty, "KRAS mutation and microsatellite instability: Two genetic markers of early tumor development that influence the prognosis of colorectal cancer," *Annals of Surgical Oncology*, vol. 17, no. 2, pp. 416–424, Feb. 2010, doi: 10.1245/s10434-009-0713-0 https://pmc.ncbi.nlm.nih.gov/articles/PMC4380015/#abstract1

[7] Y. Lao, X. Wang, Y. Yang, et al., "Characterization of genomic alterations and neoantigens and analysis of immune infiltration identified therapeutic and prognostic biomarkers in adenocarcinoma at the gastroesophageal junction," *Frontiers in Oncology*, vol. 12, Article 941868, 2022, doi: 10.3389/fonc.2022.941868 https://pubmed.ncbi.nlm.nih.gov/36439494/

[8]T. Arulraj, A. Popel, et al., "Computational Tool Developed to Predict Immunotherapy Outcomes for Patients with Metastatic Breast Cancer," *Proceedings of the National Academy of Sciences*, Oct. 2024. [Online]. Available: https://www.hopkinsmedicine.org/news/newsroom/news-releases/2024/10/computational-tool-developed-to-predict-immunotherapy-outcomes-for-patients-with-metastatic-breast-cancer. [Accessed: Mar. 11, 2025].

[9] F. Petrelli, M. Ghidini, A. Ghidini, G. Tomasello, et al., "Outcomes Following Immune Checkpoint Inhibitor Treatment of Patients With Microsatellite Instability-High Cancers: A Systematic Review and Meta-analysis," *JAMA Oncology*, vol. 6, no. 7, pp. 1068–1071, July 2020, doi: 10.1001/jamaoncol.2020.1046 https://jamanetwork.com/journals/jamaoncology/fullarticle/2765752#google_vignette

[10] P. Courtiol *et al.*, "Deep learning-based classification of mesothelioma improves prediction of patient outcome," *Nature Medicine*, vol. 25, no. 10, pp. 1519–1525, Oct. 2019, doi: 10.1038/s41591-019-0583-3. https://www.nature.com/articles/s41591-019-0583-3