FINAL PROJECT BSc COURSE MACHINE LEARNING
MSc Bioinformatics & System Biology students version

TCGA Colorectal Cancer dataset

Background

Colorectal cancer (CRC) is the third most common and second deadliest cancer worldwide, accounting for approximately 10% of cancer-related deaths in the western world. (Dekker et al., 2019)

Cancer status can highly differ between patients with the same general diagnosis for CRC and it can impact the effect of therapies. This underscores the need for new high-precision methods to predict such status and biomarkers with the final aim to provide a targeted treatment for each patient. (Li et al., 2021). You can find a more detailed description of the dataset **here**.

Data

We provide a dataset consisting of 462 colorectal cancer samples and their 33379 gene expression counts (`tcga_rna_count_data_crc.csv`). Another file containing metrics and the mutation profile for some genes of interest (`prediction_file_crc.csv`) is provided. More in detail:

- **MSI_status** (Microsatellite Instability) is the accumulation of insertion or deletion errors at microsatellite repeat sequences in cancerous cells as a result of a functional deficiency within one or more major DNA mismatch repair proteins. MSI status are three: MSI-H (Instable High), MSI-L (Instable Low), MSS (Stable). (Promega website) (Kawakami et al., 2015)
- **TBL** (Tumour Break Load) has a large impact on tumour biology and can be used as prognostic biomarker in patients with non-metastatic MSS CRC. (Lakbir et al., 2022)
- **TMB** (Tumor Mutational Burden) is a measure of the number of nonsynonymous mutations carried by tumor cells. Tumors with many mutations have a high mutational burden (high TMB). Many of the tumors with high TMB are also MSI-H.
- **fraction_genome_altered** (Nguyen et al., 2022)
- **aneuploidy_score** is calculated as the sum of altered arms, within a range of 0 to 39. (Auslander et al., 2020)
- **TP53, KRAS, BRAF, APC, TTN** are the mutation status of frequently mutated genes in CRC, but not necessaritly have a role in cancer development (e.g TTN). ("Comprehensive molecular characterization of human colon and rectal cancer", 2012)

Task

The task consists in predicting one or more labels (columns in `prediction_file_crc.csv`, e.g. MSI_status or TBL status, based on the RNA expression as input features), applying preprocessing techniques as you see fit and building model/s that suit best the task chosen. You can find more features and response variables **here**; you can use the website to get insights about the dataset and/or make your research question more challenging. For the research question, originality will be rewarded.

The general guidelines that can be found on Canvas also apply to this project.

References

Auslander, N., Wolf, Y. I., & Koonin, E. V. (2020, March). Interplay between DNA damage repair and apoptosis shapes cancer evolution through aneuploidy and microsatellite instability. *Nature Communications*, *11*(1). Retrieved from https://doi.org/10.1038/s41467-020-15094-2 doi: 10.1038/s41467-020-15094-2

Comprehensive molecular characterization of human colon and rectal cancer. (2012, July). *Nature*, *487*(7407), 330–337. Retrieved from https://doi.org/10.1038/nature11252 doi: 10.1038/nature11252

Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M., & Wallace, M. B. (2019, October). Colorectal cancer. *The Lancet*, *394*(10207), 1467–1480. Retrieved from https://doi.org/10.1016/s0140-6736(19)32319-0 doi: 10.1016/s0140-6736(19)32319-0

Kawakami, H., Zaanan, A., & Sinicrope, F. A. (2015, June). Microsatellite instability testing and its role in the management of colorectal cancer. *Current Treatment Options in Oncology*, *16*(7). Retrieved from https://doi.org/10.1007/s11864-015-0348-2 doi: 10.1007/s11864-015-0348-2

Lakbir, S., Lahoz, S., Cuatrecasas, M., Camps, J., Glas, R. A., Heringa, J., ... Fijneman, R. J. (2022, December). Tumour break load is a biologically relevant feature of genomic instability with prognostic value in colorectal cancer. *European Journal of Cancer*, *177*, 94–102. Retrieved from https://doi.org/10.1016/j.ejca.2022.09.034 doi: 10.1016/j.ejca.2022.09.034

Li, Y., Ma, Y., Wu, Z., Zeng, F., Song, B., Zhang, Y., ... Wu, M. (2021, September). Tumor mutational burden predicting the efficacy of immune checkpoint inhibitors in colorectal cancer: A systematic review and meta-analysis. *Frontiers in Immunology*, *12*. Retrieved from https://doi.org/10.3389/fimmu.2021.751407 doi: 10.3389/fimmu.2021.751407

Nguyen, B., Fong, C., Luthra, A., Smith, S. A., DiNatale, R. G., Nandakumar, S., ... Schultz, N. (2022, February). Genomic characterization of metastatic patterns from prospective clinical sequencing of 25, 000 patients. *Cell*, *185*(3), 563–575.e11. Retrieved from https://doi.org/10.1016/j.cell.2022.01.003 doi: 10.1016/j.cell.2022.01.003