Spatial Analysis of Tumor Heterogeneity Using Machine Learning Techniques

Investigators:

2022 Summer FoDOMMaT Fellow:

- Chancharik Mitra, UC Berkeley EECS Rising Sophomore

Mentors:

- Dr. Zeynep Madak-Erdogan, Associate Professor of Nutrition at UIUC
- Dr. Aiman Soliman, Research Scientist at the National Center for Supercomputing Applications (NCSA) at UIUC
- Jin Young Yoo, PhD Student at UIUC

Research Questions and Hypotheses:

Analysis of tumor samples is an important part of cancer research, and with the emergence of machine learning and computer vision techniques in recent years, they can be utilized to derive greater insight into the resistance of tumor tissue to certain therapies [2, 3]. Thus, the main question which we look to answer is: can machine learning techniques applied on spatial cancer tumor data provide insight into which regions of the tissue are resistant to therapy?

We can delve further and not just identify regions that are resistant to therapy but also fully segment tumor image data by region. The analysis of spatial data can allow us to answer many other questions with regard to the geospatial relationship of the different parts of the tissue. Other areas of exploration include understanding the location of key external structures such as blood vessels, macrophages, and hepatocytes. This is useful in understanding the microenvironmental dynamics of the tissue [4]. We can also explore whether geospatial gene expression data can give insight into the effects of the immune system on the tumor.

Background:

Data analysis and specifically machine learning techniques have been shown to be effective at analyzing spatial data of cancer tumors in the past. In one particularly promising example, one team looked at the characteristics of tumor ecology using spatial data analysis [4]. It has been posited that ecological relationships such as predation, mutualism, and parasitism exist in a tumor's microenvironment between cancer cells, regular cells, the immune system, etc. The group used methods such as SVMs and other computer vision tools to understand these ecological relationships more deeply at the level of the tumor's microenvironment [4]. With so many different features, this is not normally possible to do manually. At this scale, the genetic heterogeneity of cancer cells allows some cell types to survive under the selective pressures that mirror those at the macroecological scale [4]. Several interesting findings are as follows:

- "In our study, a high degree of co-localization between cancer and immune cells measured by this [Morista-Horn] index was found to be significantly associated with increased probability of ten-year disease-specific survival in human epidermal growth factor receptor 2-positive (Her2+) breast cancers." [4]
- The authors also mention a study where a mutualistic relationship between hypoxic and non-hypoxic cancer cells was found. At a high level, this mutualistic relationship better allows both types of cancer cells to proliferate and survive the selective pressures of the tumor microenvironment, making it a promising area of study for cancer research.

- “Cancer cells also form commensal relationships with their microenvironment. For example, cancer-associated fibroblasts (CAFs) are known to support tumor growth and progression.”[4]

The surveying of this paper for literature analysis was extremely useful in guiding the thought-process for our own line of inquiry.

We discuss another important paper briefly as well. This paper by Bergland et. al. explores the usage of spatial analysis techniques on scRNA sequence data of prostate cancer tumors. The authors develop a series of expression profiles and activity maps intended to analyze different characteristics of the tumor samples. One example of such analysis was making the distinction between healthy and diseased areas of the sample [1]. The group was able to “delineate the extent of cancer foci more accurately” than pathologist annotations [1]. These findings provide encouraging successes in the space of applying spatial analysis to sequencing data of tumor samples.

Method:

As mentioned earlier, our inquiry looks to answer more than one question, but we lay out the methodology for answering the main question of identifying regions in the tumor that are resistant to therapy using spatial analysis.

1. Individual tumor samples are prepared for analysis and passed through the Spaceranger and Loupe sequencing software (from 10X Genomics) to provide gene expression profiles of the tumor tissues. Put simply, these software take the raw data and provide locational gene expression data so we can identify which genes are expressed at which locations of the tumor.
2. This data has to be cleaned, and then exploratory data analysis has to be performed as this is a large dataset. Part of this process is specifically extracting the data that we wish to perform analysis on. At a high-level this data provides an x-y coordinate of the tissue sample paired with a barcode representing the gene expressed at this location. We will also be analyzing the raw image data of the tissue samples. In parallel to this, the data has to be labeled with the target classes representing areas resistant and not resistant to therapy.
3. We are currently at this step of the process. We take the cleaned and extracted data and perform machine learning analysis. For the tabular data, we plan to use random forest classifiers and gradient boosting techniques to classify the data. We choose these methods (1) because they work effectively (more so than other classical ML methods such as decision trees) but also because (2) they are to some extent more explainable than deep learning methods, a characteristic desirable in the medical/biological space.
4. Following this, we plan to take the image files and perform semantic segmentation. We will compare these results to the results from the previous step. We will explore several architectures for this, but for now, we are planning on using the UNET and/or DINO architectures for their usage in segmentation tasks.

Potential Benefits:

The potential benefits of this research have already been covered to some extent in the previous sections, but we summarize them here (of course, this listing is not exhaustive). The benefits include contributions to developing new cancer therapies such as immunotherapy, better understanding of cell-level dynamics of the tumor microenvironment, and a repository of results upon which future analysis can be performed.

Bibiliography:

[1] Berglund, E., Maaskola, J., Schultz, N. et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat Commun 9, 2419 (2018). https://doi.org/10.1038/s41467-018-04724-5

[2] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[4] Sidra Nawaz, Yinyin Yuan, Computational pathology: Exploring the spatial dimension of tumor ecology, Cancer Letters, Volume 380, Issue 1, 2016, Pages 296-303, ISSN 0304-3835, https://doi.org/10.1016/j.canlet.2015.11.018.