Characterization of Carcinoma Cell Culture Subpopulations using Cyclic Immunofluorescence Data & ChatGPT

Cypress Tomaneng

This case study is inspired by my work on pancreatic cancer that took place largely in 2021 & 2022 at the Developmental Oncogene Laboratory at California State University, Northridge. At the time, I was investigating the variation that existed within pancreatic cancer cell cultures with and without a particular genetic modification: the knockout via CRISPR-cas9 of retinoic acid induced 14 (RAI14). In addition, I was studying how these cells responded to being seeded in well plates with different extracellular matrix proteins, including collagen I & fibronectin. I employed cyclic immunofluorescence (CycIF), a process described by Lin et al. (2015) which involves repeated staining & imaging of cells, to gather proteomic data through fluorescence microscopy. Subsequent processing of the images, followed by data extraction & normalization that involved my own ECMAScript and Python code, allowed me to acquire workable data on the expression & subcellular localization of a panel of proteins relevant to cancer progression. A visual overview of the CycIF workflow leading up to this case study is available in this repository as `project_context.pptx`.

In the past, I was able to analyze these data using a proprietary software tool called SeqGeq. SeqGeq could perform dimensionality reduction, produce visualizations of the results, then allow for manual gating of the data for the identification and comparison of subpopulations of cancer cells. However, as I favor tools with less restrictive (and less expensive) licenses, I took this course as an opportunity to attempt to develop a more flexible and affordable solution for analysis of CycIF data. I prompted ChatGPT to do the same initial tasks, including t-SNE dimentionality reduction & visualization, then attempted to explore the data with further assistance from ChatGPT for data clustering and comparison of subpopulations. I did so in two conversations. The first of these, which I have designated conversation A, involved me describing the experiment, data format and file organization in short prompts, along with requests for code to perform the desired tasks one step at a time. The second of these, which I have designated conversation B, involved me condensing my experiment description and much of my previous prompts into a single prompt. I chose to end each of these conversations when ChatGPT began repeatedly misdiagnosing a bug that I was unable to correct.

Within this repository, the first conversation is documented in the `conversation_a` directory, and the second is documented in the `conversation_b` directory. Within each directory, all prompts & responses leading up to a given code snippet are logged in `.txt` files with numbers assigned in the order they occurred. Each code script or snippet is stored as a `.R` file within the same directory and numbered the same as the corresponding piece of conversation with ChatGPT. This parent directory also contains a minimally edited version of what I consider to be the most successful/impressive script generated by ChatGPT (“`conversation_a_code_13_edited.R`”), along with the image it produced (“`example_t-SNE_plot_with_clustering.PNG`”).