**Artificial Intelligence-Driven Automation of Flow Cytometry Gating**

Gabriella Rivera

Applied Data Science Master’s Program

Shiley Marcos School of Engineering / University of San Diego [gabriellarivera@sandiego.edu](mailto:gabriellarivera@sandiego.edu)

John Vincent Deniega

Applied Data Science Master’s Program  
Shiley Marcos School of Engineering / University of San Diego   
 [jdeniega@sandiego.edu](mailto:jdeniega@sandiego.edu)

**ABSTRACT**

Flow cytometry is a biochemical process that measures the physical and chemical characteristic of cells in a liquid suspension. This method enables the identification and classification of various cellular populations, such as lymphocytes, monocytes, and granulocytes – from Peripheral Blood Mononuclear Cell (PBMC) samples. Clustering algorithms aim to objectively simplify the classification accuracy of a given population's cellular type to further immunological and clinical diagnostic purposes. By exploring algorithms like K-means, Agglomerative, and Gaussian Mixture Modeling, the dataset containing metrics on front and side scatter areas of cell scans, fluorescence on markers, and other dimensions can potentially lead to automatic gating and bring forth insights on cell population characteristics, resulting in increased analytical throughput and increased rate of medicinal and pharmacological breakthroughs.

1 Introduction

According to Brestoff and Frater (2022), flow cytometry is a cost-prohibitive cellular identification and pharmacological discovery process that both requires exceptional capital investment in infrastructure and is inherently difficult in adopting and implementing the latest techniques and technologies in the field. Open-source and low-cost options may serve as a viable bridge technology between capital-intensive investment cycles for laboratories to be able to continue evolving and increasing their analytical throughput without relying on the next hardware upgrade. By using existing mathematically-driven principles inherent to artificial intelligence, biochemists may be able to reduce the analytical inputs required to perform routine cellular classification and clustering of PBMCs. This labor-intensive process ultimately serves to evaluate the efficacy of experimental groups related to clinical trials that are required in the development of new life-saving medicines.

2 Background

Flow cytometry can be a capital-intensive process that requires significant investments in laboratory-grade biomedical equipment, dedicated graphics processing and tensor processing units, expansive random-access memory, and proprietary analytical software licenses. With Flow Cytometry Standard (FCS) files readily available on public repositories and by leveraging open-source and permissive license packages such as Scikit-Learn and Matplotlib to perform computational transformations to FCS data, we aim to discover cost-effective alternatives to expensive enterprise software licenses that perform flow cytometry analysis, which may result in significant reduction in the barriers to entry in biochemical flow cytometry.

Because cellular populations related to these FCS files number in the millions of records across multiple laboratory readings, this project will place heavy emphasis on dimensionality reduction in order to meet the constraints of being both cost-effective and hardware resource-efficient. Accomplishing such a feat would result in independent biochemical scientists to perform analyses without relying on exceptionally powerful computing hardware resources or costly proprietary enterprise-level software licenses.

2.1 Problem Identification and Motivation

As of this publication, flow cytometry gating is a manual process that involves a highly-trained biochemist to process and analyze the results of optical scans of cellular assays that may be further augmented by fluorescent substrates. Because of the complex and highly-dimensional nature of the data, these scientists rely on a best-practices approach based on their own respective processes and frameworks. Because of the potential variability of these processes and frameworks, the resulting findings from interpreting scan results is dependent on both the breadth and depth of methods of a given supervising scientist, thus resulting in both an increase in cost of analysis due to human error and omission as well as a reduction in consistency of results.

2.2 Definition of Objectives

The research team aims to utilize open-source and publicly-available resources from well-known algorithms known in data science to include principal component analysis, t-distributed stochastic neighbor embedding, and unsupervised clustering machine learning methods as well as FCS data hosted by FlowRepository (2020). Once data is cleaned for noise from scan data, the team aims to train models or machine-learning applications that have potential for value-added analysis relative to that of a typical human biochemist. Upon evaluation, success is generally defined when automated analysis reaches parity with a human analyst of at least 90% classification accuracy of PBMCs toward their respective dendritic cellular type on an unseen test set containing FCS scan data. In the event that this evaluation criterion is not met, further justification would have to be provided whether the measured degree of accuracy is acceptable relative to the speed of analyses.

3 Literature Review (related works)

Since 2016, a number of academic threads have been studied involving the advancement in flow cytometry, the iteration of methodologies when incorporating machine learning applications on FCS data, as well as different strategies in how to potentially automate the classification of cellular groups. By 2024, Ng et al. (2024) demonstrate maturity over an eight-year period that transitions the focus of academic research from the “what” normally seen in earlier works into the “how” with respect to interdisciplinary guidelines as well as quality control and assurance of future deployment of artificial intelligence in flow cytometry (p. 228).

3.1  FlowAI: Automatic and interactive anomaly discerning tools for flow cytometry data

FlowAI is a software package for the statistical computing language R, which Monaco et al. (2016) developed as a means to both clean FCS files from anomalies and to assess the resulting quality of the cleaned data normalized by the flow rate of a given reading. When flow rate abruptly changes during a scan, the readings may exhibit data inconsistencies. These data inconsistencies are considered anomalous and are discarded from the dataset. Using time-series analysis, the resulting dataset is broken into trend and cyclical components before being normalized by penalization function measuring absolute deviation of a data point from the median. Monaco et al. (2016) place an emphasis on data quality and anomaly handling, which are crucial considerations to flow cytometry, however they do not address the next step in automatic gating of cellular types, which is the focus of our research.

3.2  An open-source solution for advanced imaging flow cytometry data analysis using machine learning

Hennig et al. (2017) identify the challenges associated with the manual and subjective nature of flow cytometry, resulting in inconsistent in analysis. The given solution is to utilize open-source software, CellProfiler, to use raw image files to identify cell types from a flow cytometer image. Our research shares the open-source idea of being able to leverage existing machine learning algorithms to automatically classify these cell types. Contrasting the team of Hennig et al. (2017) on classification differs greatly in their use of visual image data as the basis for classification rather the numerical scan data from fluorescent biological marker excitation that is central to our approach (p. 202).

3.3  Comprehensive phenotyping of human dendritic cells and monocytes

Mair and Liechti (2020) identify the potential benefits in using biological markers to identify the phenotypes specific to dendritic cells and monocytes for cellular classification. This particular research focuses on a potentially more significant subset of biological markers and lineages that aim to more precisely identify different cellular categories as a result of their fluorescence excitation scan data. This work serves as the source data of our project which uses Python-based machine learning packages for automatic gating. A similar methodology was employed by Hennig et al. (2017) who instead synthesized with visual imagery data with the open-source software, CellProfiler.

3.4  Application of machine learning for cytometry data

Hu et al. (2022) acknowledge the complex challenge of highly-dimensional flow cytometry data and the potential for existing machine learning software packages to perform analysis on this type of data. This particular team first focuses on dimensionality reduction by means including Principal Component Analysis and stochastic methods, unsupervised and supervised machine learning methods to predict resulting clinical outcomes such as healthy populations versus diseased populations (p. 2). Our project aims to build on this research with greater training and tuning toward existing biological knowledge cross-validated across different FCS file scan results.

3.5  Recommendations for using artificial intelligence in clinical flow cytometry

Most recently, Ng et al. (2024) focuses on a more interdisciplinary approach to using artificial intelligence in flow cytometry with unique considerations for clinical risk management, quality control and assurance, and computational efficiency. This requires extensive consideration as to the narrative annotations required for clinical implementation. Though the article is comprehensive across multiple sectors related to flow cytometry and the technical and regulatory nuances required when applying artificial intelligence, it only provides general recommendations and guidance for future scientist who wish to leverage this new technology. Relative to our existing work, our research team aims to apply these general recommendations and implement them in an open-source and demonstrable product for flow cytometry automatic gating.

4 Methodology

The flow cytometry dataset was acquired from FlowRepository.

4.1 Data Extraction and Format Conversion

4.2 Data Feature Selection

4.3 Exploratory Data Analysis

4.4 Data Cleaning

4.5 Dimension Reduction

4.5.1 PCA

4.5.2 T-SNE

4.5.3 UMAP

**References**

Beckman Coulter. (2022). *Automatic gating - Beckman Coulter*. Flow Cytometry. <https://www.beckman.com/flow-cytometry/software/cytobank-premium/learning-center/automatic-gating>

Brestoff, J. R., & Frater, J. L. (2022). Contemporary challenges in clinical flow cytometry: Small samples, big data, little time. *Journal of Applied Laboratory Medicine*, *7*(22), 931-944. <https://doi.org/10.1093/jalm/jfab176>

Cossarizza, A., Chang, H. D., Radbruch, A., Acs, A., Adam, D., Adam-Klages, S., Agace, W. W., Aghaeepour, N., Akdis, M., Allez, M., Almeida, L. N., Alvisi, G., Anderson, G., Andrä, I., Annunziato, F., Anselmo, A., Bacher, P., Baldari, C. T., Bari, S., Barnaba, V., … Zychlinsky, A. (2019). Guidelines for the use of flow cytometry and cell sorting in immunological studies (second edition). *European journal of immunology*, *49*(10), 1457–1973. <https://doi.org/10.1002/eji.201970107>

FlowRepository. (2020). FlowRepository ID FR-FCM-Z32U. Retrieved from <http://flowrepository.org/experiments/3166/download_ziped_files>

Hennig, H., Rees, P., Blasi, T., Kamentsky, L., Hung, J., Dao, D., Carpenter, A. E., & Filby, A. (2016). An open-source solution for advanced imaging flow cytometry data analysis using machine learning. *Methods*, 112(2017), 201–210. <http://dx.doi.org/10.1016/j.ymeth.2016.08.018>

Hu, Z., Bhattacharya, S., & Butte, A. J. (2022). Application of machine learning for cytometry data. *Frontiers in immunology,* 12, 787574. <https://doi.org/10.3389/fimmu.2021.787574>

Lee, J. A., Spidlen, J., Boyce, K., Cai, J., Crosbie, N., Dalphin, M., Furlong, J., Gasparetto, M., Goldberg, M., Goralczyk, E. M., Hyun, B., Jansen, K., Kollmann, T., Kong, M., Leif, R., McWeeney, S., Moloshok, T. D., Moore, W., Nolan, G., Nolan, J., … Brinkman, R. R. (2008). MIFlowCyt: the minimum information about a Flow Cytometry Experiment. *Cytometry. Part A: the journal of the International Society for Analytical Cytology*, 73(10), 926–930. <https://doi.org/10.1002/cyto.a.20623>

Maecker, H. T., McCoy, J. P., & Nussenblatt, R. (2012). Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews Immunology,* 12(3), 191–200. <https://doi.org/10.1038/nri3158>

Mair, F., & Leichti, T. (2020). Comprehensive Phenotyping of Human Dendritic Cells and Monocytes. *Journal of Quantitative Cell Science*, 99(3), 231–242. <https://doi.org/10.1002/cyto.a.24269>

Monaco, G., Chen, H., Poidinger, M., Chen, J., Magalhaes, J., & Larbi, A. (2016). FlowAI: Automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics*, 32(16), 2473–2480. <https://doi.org/10.1093/bioinformatics/btw191>

Ng, D. P., Simonson, P. D., Tarnok, A., Lucas, F., Kern, W., Rolf, N., Bogdanoski, G., Green, C., Brinkman, R. R., & Czechowska, K. (2024). Recommendations for using artificial intelligence in clinical flow cytometry. *Cytometry part b: Clinical cytometry*, 106(4), 228–238. <https://doi.org/10.1002/cyto.b.22166>

Spidlen J, Breuer K, Rosenberg C, Kotecha N and Brinkman RR. (2012). FlowRepository - A Resource of Annotated Flow Cytometry Datasets Associated with Peer-reviewed Publications. *Cytometry A*. 81(9), 727-31. <https://doi.org/10.1002/cyto.a.22106>