Label-Free Classification for Flow Cytometry

Mohammad Tanhaemami[1], Elaheh Alizadeh[1], Claire Sanders [2], Babette Marrone [2] Brian Munsky[1,*]

***Abstract*—In flow cytometry, specific biochemical labels are not always available; they can be costly; and they can disrupt natural cell behavior. Label-free classification strategies are needed to correct these issues. Unfortunately, label-free strategies may be difficult to learn when applied labels or other modifications in training data inadvertently modify intrinsic cell properties. We develop a new approach based upon population statistics and machine learning to integrate labeled and unlabeled training data and to identify models for quantitatively accurate label-free classification. We apply our approach to make label-free measurements of lipid content in microalgae cells.**

***Keywords*—Single cells, flow cytometry, machine learning.**

I. Introduction

FLOW cytometry (FCM) is an essential single-cell measurement technique that uses biochemical labels to mark cell features, quantify cell properties, or sort cell populations. However, use of biochemical labels can interfere with FCM inferences or disrupt cellular behaviors [1]. From massive datasets generated by FCM, statistical information on intrinsic cell populations can be employed with machine learning (ML) tools to develop label-free strategies [2]. In this work, we combine ML approaches with single-cell fluctuation fingerprint analyses [3] and genetic algorithm-based feature selection to find optimized label-free classification for lipid accumulation in algal cells.

II. Results

We monitored *Picochlorum* microalgae for 16 days following nitrogen starvation, and we created two identical population samples at each of 13 timepoints. We stained one set of samples with boron-dipyrromethene (BODIPY) and left the other unstained. Using an ACCURI^{TM} flow cytometer, we measured 3000 cells for each sample at each timepoint, and we recorded 12 features per cell, including the 488nm excited 530/30nm fluorescence channel corresponding to the BODIPY dye. We then sought to predict the BODIPY signal intensities using the other measured features (e.g., autofluorescence and side scatter at other fluorescence wavelengths). Preliminary regression analysis suggests a strong classification for the training data (Fig. 1a), but this regression model fails to correctly estimate lipid content in the absence of labels, i.e., in testing phase (Fig. 1b). We quantify the accuracy of predictions for unlabeled cells using the Kolmogorov-Smirnov (KS) distance between measured and predicted lipid distributions.

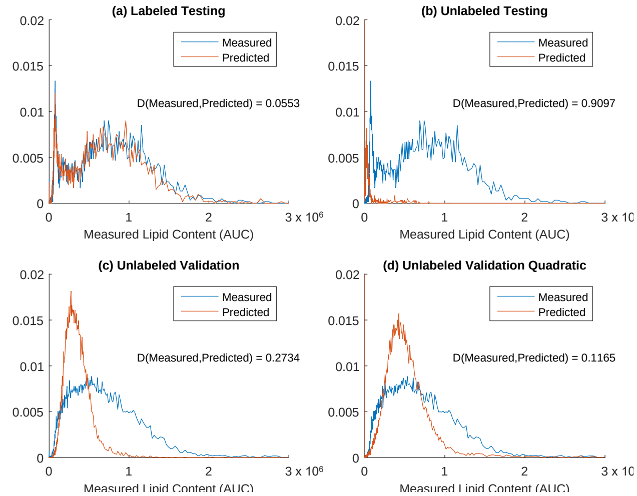

Fig. 1. Label-free classification for single-cell populations. Measured (red) and predicted (blue) distributions of cell lipid content in arbitrary units of concentration (AUC). (a) Labeled testing data and (b) unlabeled testing data using linear regression. (c) Unlabeled validation data using the genetic algorithm for feature set reduction. (d) Unlabeled validation data using quadratic features and the genetic algorithm. Kolmogorov-Smirnov distance between distributions is shown for each case.

Next, we use a genetic algorithm and an iterative training/testing strategy to identify a reduced feature set that minimizes the KS distance to enhance label-free classification. Expansion of the single-cell feature sets to include quadratic feature combinations followed by reduction with the genetic algorithm results in substantially improved label-free predictions for the lipid content (Fig. 1c). We verify that our new model could be used for label-free estimation of single-cell lipid content.

III. Conclusions

We apply mathematical models, machine learning and genetic algorithms to circumvent the need for biochemical labels. Because such labels can be expensive or disruptive to natural cell behavior, the use of computational models to replace chemical biomarkers can open new avenues for single-cell research. The key to our approach is to iterate between labeled and unlabeled data and to carefully remove unnecessary or misleading features. In future work, we will use these classification strategies to sort single cells into different subpopulations without the disruptions associated with biochemical markers.

References

[1] Gossett, et al, (2010), *Analytical and Bioanalytical Chemistry*, **397**: 8, 3249-3267

[2] Saeys, et al, (2016), *Nature Reviews Immunology*, **16**: 7, 449-462

[3] Munsky, et al, (2012), *Science*, **336**: 6078, 183-187

Acknowledgements: This work was funded in part by NIGMS-R25GM105608.

[1] Chemical and Biological Engineering, Colorado State University; Fort Collins, CO, USA.

[2] Bioscience Division, Los Alamos National Laboratory.

[*] Correspondence: munsky@colostate.edu