USING GAUSSIAN MIXTURE MODELS AND MACHINE LEARNING TO PREDICT DONOR-DEPENDENT MEGAKARYOCYTIC CELL GROWTH AND DIFFERENTIATION POTENTIAL EX VIVO

William M. Miller, Chemical and Biological Engineering, Northwestern University, Evanston, IL
wmmiller@northwestern.edu
Jia J. Wu, Darryl A. Abbott, Neda Bagheri; Chemical and Biological Engineering, Northwestern University
Meryem K. Terzioglu, Dolores Mahmud, Nadim Mahmud; Hematology/Oncology, University of IL at Chicago

Key Words: culture-derived platelets, flow cytometry, cell therapies, time-course data, cell-surface markers

The ability to analyze single cells via flow cytometry has resulted in a wide range of biological and medical applications. Currently, there is no established framework to compare and interpret time-series flow cytometry data for cell engineering applications. Manual analysis of temporal trends is time-consuming and subjective for large-scale datasets. We resolved this bottleneck by developing TEmporal Gaussian Mixture models (TEGM), an unbiased computational strategy to quantify and predict temporal trends of developing cell subpopulations indicative of cellular phenotype.

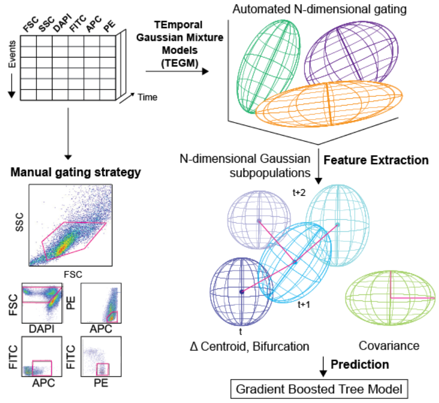

Figure 1 – Gaussian mixture models characterize each population subset in MK culture time-course data.

TEGM applies Gaussian mixture models and gradient boosted trees for cell engineering applications. TEGM enables the extraction of subtle features, such as the dispersion and rate of change of surface marker expression for each subpopulation over time. These critical, yet hard-to-discern, features are fed into machine-learning algorithms that predict underlying cell classes. Our framework can be flexibly applied to conventional flow cytometry sampling schemes, and allows for faster and more consistent processing of time-series flow cytometry data.

As a proof-of-concept, we applied our method to the analysis of *ex vivo* megakaryocytic (Mk) differentiation and maturation of hematopoietic cells from donors with greatly varying potential to generate CD41$^+$CD42$^+$ mature Mk cells. We illustrate the major steps of the computational approach by predicting peak %CD41$^+$CD42$^+$ MK maturation of CD34$^+$-selected umbilical cord blood (CB) cells from 16 independent donors (Figure 1). Cells were cultured over a 19-day multi-phase differentiation culture, consisting of a pre-expansion phase and a differentiation phase. The novel dataset comprised 720 measurements from 80 perturbations of 16 individual donors, with 9 time-point measurements sampled every 2-5 days for each donor. We constructed an automated gating strategy to extract surface marker expression of various clusters of DAPI^{low}CD41$^+$ Mk cells. Notably, we demonstrate that estimation of the %CD34$^+$ and %CD42$^+$ cells was within 1% of manual gating estimates, thus illustrating the consistency and accuracy of the technique. A gradient boosted tree model was trained using an explanatory matrix describing early characteristics and tested to predict peak CD41$^+$/CD42$^+$ marker expression. We then performed feature extraction for each flow cytometry time-course dataset on several descriptors, such as growth rate, viability, production, percentage positivity of each surface marker, covariance of mean fluorescence intensity, rate of change, and bifurcation of each subpopulation. A gradient boosted tree model was trained using an explanatory matrix describing early characteristics (Day 0 to Day 9) and tested to predict peak CD41$^+$CD42$^+$ marker expression, which typically occurs on Day 14 to Day 17.

Overall, we identify several influential early culture factors that are predictive of peak %CD42$^+$ expression. We show that %CD41$^+$ on Day 5 and Day 7 is highly predictive, while cell viability and %CD34$^+$ are comparatively less predictive of peak %CD42$^+$. We are able to identify the best performing cultures with high sensitivity and specificity (AUROC = 0.92, where 1 denotes perfect accuracy). Predicted and actual CD41$^+$CD42$^+$ responses are highly correlated using three independently selected partitions of test/training sets of our data (Figure 2; p = 7.4e-09, R = 0.87). Identifying CB units with high and low MK potential early in the 19-day culture process can save expensive resources and time, and provides the potential to intervene during the culture process.

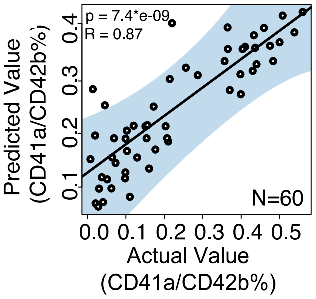

Figure 2 – TEGM with gradient boosted tree model predicts peak %CD41a$^+$/ CD42b$^+$ of Mk cultures.