Isolating single cell types from co-culture flow cytometry experiments using n-dimensional gating for CAR T-based cancer immunotherapy

Victor Tieu[1] (vtieu@stanford.edu)

[1]Department of Bioengineering, Stanford University, Stanford, CA

Background

Flow cytometry is a method of single cell analysis where cells are encased in individual microfluidic droplets and run through a series of lasers. The light that is emitted by each cell is collected and processed as a cell signature—these features can be represented on a 2D dot plot, and "gates" can be drawn manually (and somewhat arbitrarily) to partition off different populations of interest.

Gating flow data with mixed cells is challenging, since unique cell "markers" must be present in order to differentiate cell types. In co-culture experiments such as CAR T cell + tumor cell functional assays, the levels of these markers often change due to cell-cell interactions (i.e. cells no longer "look" like the unmixed populations). Therefore, an unsupervised learning algorithm is appropriate to identify clusters within flow data that represent a single cell type. In this project, I evaluated the ability of k-means clustering and Gaussian mixture models, in combination with PCA, to isolate single cell types, and trained a semi-supervised EM algorithm to do so.

Methods and implementation

I collected flow data from an experiment with mixed primary human CAR T cells and K562 leukemia cells. FCS files were converted to CSV using an open-source script provided by GenePattern and the Broad Institute at MIT. Each row in the design matrix represents a single cell, and each column represents one of seven "channel" features (FSC-A/H, SSC-A, APC, mCherry, EGFP, AF405) from the raw input data. Derived features include PCA projections of the input for visualization and clustering due to correlation between channels. Unmixed cells were used as the ground truth-labeled dataset. Simple data pre-processing included shifting all values by the min, taking the log10 of color channels (since the distribution of fluorescence intensity is log10-normal), and setting mean and variance to 0 and 1.

After projecting to a lower-dimensional subspace using PCA:

$$u^T \left(\frac{1}{m} \sum_{i=1}^{m} x^{(i)} x^{(i)T} \right) u.$$

clusters were assigned by finding centroids (k-means):

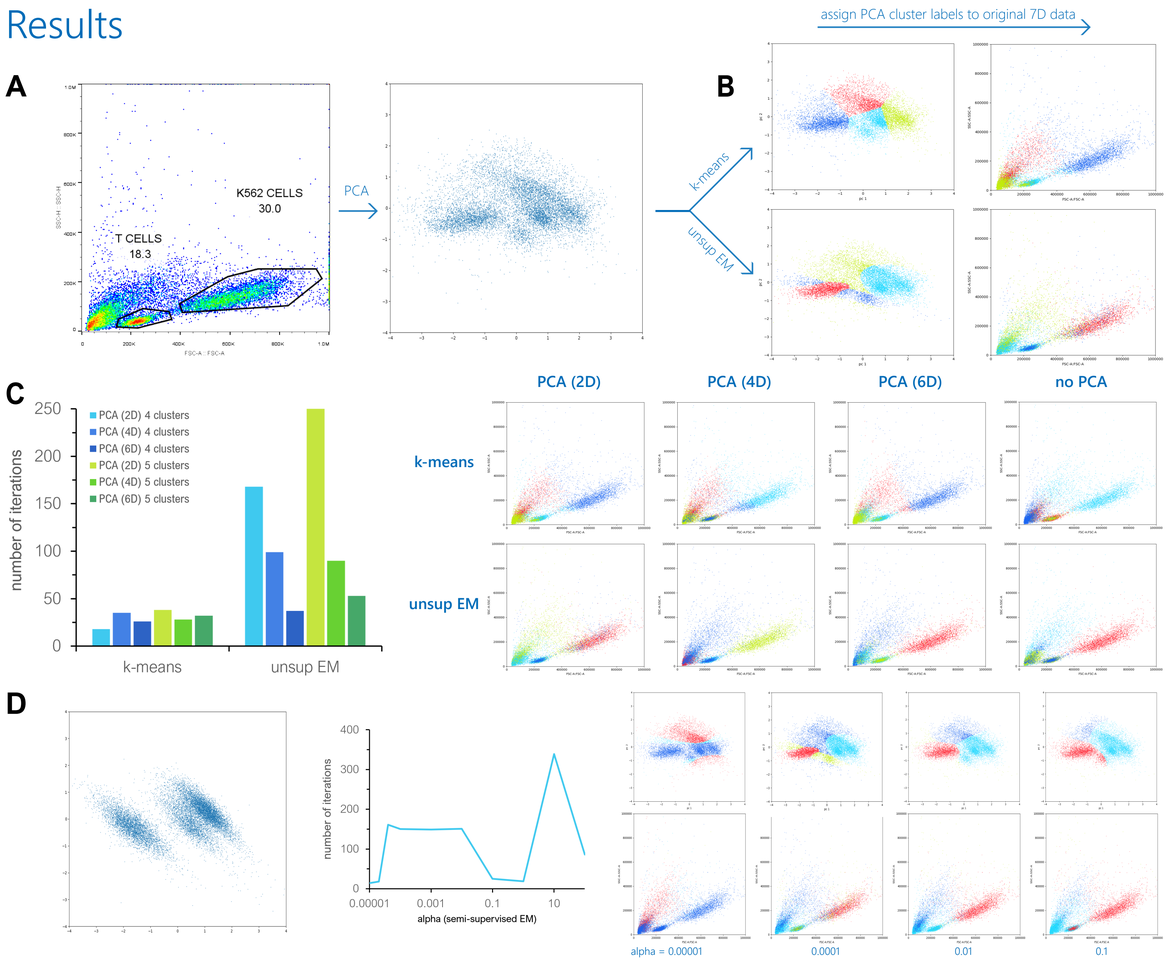
$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2. \qquad \mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

or by fitting data to a mixture of Gaussians and maximizing the log-likelihood through the EM algorithm:

$$Q_i^{(t)}(z^{(i)}) := p(z^{(i)}|x^{(i)}; \theta^{(t)})$$

$$\theta^{(t+1)} := \arg\max_\theta \left[\sum_{i=1}^{m} \left(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i^{(t)}(z^{(i)})} \right) + \alpha \left(\sum_{i=1}^{\tilde{m}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta) \right) \right]$$

Results



A. Raw 7-dimensional data for the mixture of T cells and target cells (left, plotted in two-dimensions as forward- v. side-scatter) is projected onto a two-dimensional subspace by PCA (right). Each dot represents a single cell (feature vector). The manual gating strategy is shown on the original dot plot for qualitative comparison with clustering results.

B. Clustering of the PCA projections is shown for k-means (top left) and unsupervised EM (bottom left) with k=4 clusters. The resulting cluster labels are reassigned to the non-projected raw data for both algorithms (top and bottom right, respectively).

C. Both models were run using PCA projections with 2, 4, and 6 principal components for k=4 or k=5 clusters. Number of iterations until convergence is shown for all conditions run (left). Clustering results are shown for all k=4 conditions (right).

D. Semi-supervised EM was run using unmixed cells as labeled controls. The labeled dataset was projected into the 2D PCA subspace with the same eigenvectors generated from the unlabeled set (left). The number of iterations until convergence is plotted against varying alpha coefficients for the labeled set (middle). Clustering results are shown on the right for different alpha values (cluster number k=4).

Discussion

Qualitatively, unsupervised EM outperforms k-means clustering in comparison to the manual gates set for T cells and K562 cells. However, for k-means, the algorithm converges relatively quickly and with little correlation to principal component number or cluster number. For unsupervised EM, smaller PCA subspace dimension and greater cluster number seem to slow algorithm convergence while improving clustering quality.

To improve convergence rate and stability, I tried to introduce a supervised term into the EM algorithm using unmixed positive cells. However, semi-supervision doesn't really work well in this case, since interactions between mixed cell populations changes how they compare to the ground truth in the PCA subspace (note the cluster difference between plots in A and D). Trying out a range of alpha values results in varied clustering quality and convergence rate (plot E). Interestingly, alpha = 0.0004 resulted in the same clustering as the unsupervised case, which leads me to believe that giving the algorithm a very small weighting toward the unmixed PCA clustering helps stability/convergence rate, but also that clustering can be successful (though unstable to noise) when completely unsupervised.

Future work

Since the feature vectors change with time in co-culture, an immediate next step would be to improve clustering of the same cell populations as the various features of each cluster deviate from the ground truth label. To do this, it would be interesting to explore the inclusion of a skewing factor or exponential factor within the GMM, since many of these distributions are actually skew-normal distributions[1]. Furthermore, it might be interesting to try kernelizing the GMM model to allow for infinite-dimensional Gaussian fits to the data. Tuning various hyperparameters such as cluster number k, alpha value in the semi-supervised EM, and PCA subspace dimension might also result in better clustering in the current models. Finally, it would be useful to test the current model with a dataset containing flow data of different cell types from the ones that the algorithm was trained on to see whether or not the fit is generalizable to other cell types (e.g. T cells against melanoma).

Acknowledgments and references

This project is a class assignment for CS229: Machine Learning at Stanford University. Funding and materials for flow cytometry and co-culture experiments were provided by the National Science Foundation (NSF-GRFP) and the Stanley Qi Lab at Stanford. Equations for each model were sourced from the lecture notes posted online at cs229.stanford.edu.

[1]https://doi.org/10.1016/0022-1759(85)90045-6