SWIFT analysis of FlowCAP challenges

Tim Mosmann
Gaurav Sharma
Jonathan Rebhahn
Iftekhar Naim
Jason Weaver
Suprakash Datta
James Cavenaugh

NIH: Rochester Human Immunology Center
Automated detection of rare, cytokine-producing T cells in large, high-dimensional flow cytometry datasets

Automated multivariate clustering is better:
- Reproducible, objective
- Large clinical trials
- Simultaneous analysis of many dimensions
- Discovery

Challenges: Many cells, many dimensions
- >1 million cells
- 20 variables, 16 fluorescence and 4 scatter channels

Our goal: automatically identify and compare rare cytokine-secreting cell populations in large samples

Iftekhar Naim

Gaurav Sharma
Three steps in SWIFT to adjust cluster numbers and identify rare populations

Initial populations:
- May be skewed;
- May overlap;
- May include a high dynamic range.

1: EM fitting
The EM algorithm fits the data to a specified number of Gaussians, by weighted, iterative sampling. Large asymmetric peaks may be split into multiple Gaussians, but very small peaks may not be separated.

2: Splitting
Each cluster from Step 1 is tested by LDA for multiple modes in all combinations of dimensions. Clusters are split if necessary (using EM), until all are unimodal.

3: Merging
All cluster pairs are tested for overlap, and merged if the resulting cluster is unimodal in all dimensions. Agglomerative merging prevents over-merging due to ‘bridging’ Gaussians.

The three-step procedure in SWIFT addresses several clustering challenges:
- Weighted sampling in step 1 scales to very large, high-dimensional datasets (e.g. 10 million cells, 20 dimensions);
- Splitting in step 2 identifies very rare populations;
- Merging in step 3 allows SWIFT to describe non-Gaussian clusters;
- Combined splitting and merging converges on a stable number of clusters over a wide range of input numbers;
- Soft clustering describes overlapping populations more effectively than gating.

One-dimensional examples are shown for simplicity – in reality SWIFT clusters simultaneously in all dimensions.
A PBMC sample (0.1 million cells, 7 parameters) was clustered with varying input numbers of clusters for the initial EM step. Cluster numbers were increased after the splitting step, and reduced after the merging step.

SWIFT is self-adjusting – after splitting and merging, similar output cluster numbers are obtained.

Variability between clustering runs: stochastic nature of the EM initialization, and genuine biological ambiguity resulting in alternative cluster solutions.
Comparing samples: co-clustering and templates

Clustering of flow data has multiple valid solutions, so comparisons between independently-clustered samples are difficult.

- **Solution:**
  - Merge files electronically, cluster as a single sample. This rigorously compares samples, e.g. positive and negative controls, in the same clusters.

- **Similar strategy:**
  - Produce a *cluster template* from one sample (or a consensus sample)
  - Assign cells in additional samples to this template.

Small populations (e.g. 3 cells) in negative controls cannot be clustered!
A PBMC sample from subject P, replicate 1 was clustered, generating a cluster template. Cells in additional samples were then assigned to this template. We compared assignment to the same replicate; two replicates of the same subject; pairs of different subjects; or two replicates from a second subject.
Robustness of SWIFT analysis – cells/cluster

- Three subjects, eight blood samples, two influenza stimulations. 48 files.
- Single SWIFT clustering, assign all files to this template (403 clusters).
- Determine correlation coefficients between all possible pairs of samples.
Robustness of SWIFT analysis – fluorescence intensity

- Correlations were measured between the CLUSTER MEDIANS of the fluorescence (CD3) of all pairs of samples.
Visualizing clusters: Gating on cluster medians

- After clustering, each cell is assigned two sets of values – the original, private fluorescence intensity in each channel, and the median values of its cluster.
- Using normal flow cytometry analysis programs, the results can be visualized as individual cells, or as clusters.
- Conventional gating can then be used to identify intact clusters.
Activated CD4 T cell clusters found by SWIFT

Triplicate samples of human PBMC, about 1.5 million cells each, were stimulated with Influenza peptides, or left unstimulated. Activated CD4 T cell clusters were identified by SWIFT.
Can SWIFT detect really small populations?

Concatenate 18 files, weak responses and negative controls. Cluster in SWIFT.

27.1 million events
1085 clusters
3 small clusters

19 parameters
29 CD3+ CD4+ CD45RA- CD69+ clusters

Cluster 1
25 cells

Cluster 2
20 cells

Cluster 3
21 cells

Sensitivity: better than one part per million
Correlation of manual and automated analysis

- Eight PBMC samples each from two subjects were stimulated with the polyclonal activator SEB, influenza peptides, or no antigen, and analyzed by intracellular cytokine staining.

- The Flow Cytometry files were analyzed independently by two manual operators, and also by two sets of clustering and template assigning in SWIFT. Total CD4 T cell numbers expressing IFNg and TNFa are compared.
Challenge 1A

- Challenge: identify the cells belonging to two rare populations, as described by the manual gating in the training set.

- Our Strategy: Cluster a concatenate of samples using SWIFT (three runs), and assign all samples to the cluster templates.

- Identify the clusters (of rare cells) containing the highest numbers of the two populations tagged in the training set, and report the cells in the same clusters in the test set.
Challenge 1A

- Training Cells
- Training cells in SWIFT cluster
- SWIFT cluster cells
Challenge 1A

- Problems/challenges/discrepancies:
  - Multi-dimensional gating can often identify slightly larger populations that manual bivariate gating, resulting in apparent false positives.
  - Multi-dimensional gating can often exclude contaminating populations more effectively, resulting in apparent false negatives.
  - Model-based clustering will not give a good match to manual gating of the edge of a larger population.
Challenge 3

- Challenge: Classify samples, stimulated or not stimulated with HIV antigens, into pre- and post-vaccination samples.

- Expectation: Changes in small cytokine-secreting populations would be key alterations.

- Strategy:
  - Normalize data (simple channel-specific scaling).
  - Use SWIFT to cluster a concatenate of all POL samples.
  - Assign all samples to this template.
  - SVM (Matlab) to identify features that distinguish visits in training set.
  - Assign test set.

- Small cytokine-secreting cell populations (in response to POL) were not the discriminating populations.
## Acknowledgements

- **Influenza responses:**
  - Jason Weaver
  - EunHyung Lee
  - David Roumanes
  - Martin Zand
  - Xi Li
  - Hulin Wu
  - Nan Deng
  - John Treanor

- **Amphiregulin**
  - Yilin Qi
  - Steve Georas

- **Flow Cytometry analysis:**
  - Iftekhar Naim
  - Jason Weaver
  - Gaurav Sharma
  - Sally Quataert
  - Suprakash Datta
  - Jonathan Rebhahn
  - James Cavenaugh

- Rochester Human Immunology Center, CEIRS/New York Influenza Center of Excellence, Center for Biodefense Immune Modeling, American Asthma Foundation