flowMerge: Merging Mixture Components to Identify Distinct Cell Populations in Flow Cytometry

Greg Finak, PhD

R. Gottardo Laboratory,
Vaccine and Infectious Disease Division
Fred Hutchinson Cancer Research Center

flowCAP 2010 Summit. September 21, 2010
Outline

1 Introduction
   - Goals of Automated Gating
   - Challenges for Automated Gating

2 The flowClust and flowMerge Algorithms
   - The flowMerge Algorithm
   - Shortfalls of flowClust / flowMerge

3 flowCAP
   - Algorithm Settings and Gating Strategy
   - Our Take-home lessons
   - Future Improvements to flowClust / flowMerge
   - Acknowledgements
Goals of Automated Gating

In an Ideal World:
Identify the same populations that a human expert can identify... as well as those they can’t.

Specifically
- Identify biologically relevant cell populations.
- Classify events into (one or more) of the identified cell populations.
- Do it accurately. (relative to some standard)
- Do it quickly (or at least faster than the human expert).

Many approaches, both parametric and non-parametric.
Characteristics of FCM Data

Globally FCM Data is Well Represented by a Mixture of Distributions. However

Cell populations in FCM data tend to be noisy, asymmetric, overlapping and not always well resolved by existing markers. Not all populations in an experiment are of interest to the question at hand.

From a modelling perspective.

- The distributions of individual cell populations are not "nice"
- Noisy, non-gaussian, asymmetric, have non–constant variance.
- Gating strategies depend on the data.
Quick Intro to Mixture Models I

Mixture Model

Model a complicated distribution using a weighted combination of "simpler" distributions.

\[ f_0(y) = \sum_{g=1}^{G} \pi_g f_g(y | \theta) \]

\(f_g(\cdot)\)'s can be any distributions. In practice:

- Multivariate Gaussian
- Multivariate–t (flowClust)
- Multivariate–t with Box–Cox transformation (flowClust)
- Skewed Multivariate–t (FLAME)
Quick Intro to Mixture Models II
Quick Intro to Mixture Models III

Gaussian Mixtures
- Spherical or ellipsoidal covariance

t Distribution
- Robust to outliers

Box–Cox Transformation
- Allows for asymmetry
The Box–Cox Transformation

- Flow cytometry data is usually transformed prior to gating.
- arcsinh, log, logicle, Box–Cox
- Individual populations can still be skewed.

Generalized Box–Cox Transform

\[ x = \begin{cases} 
\frac{\text{sgn}(y)|y|^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\
\log(y) & \text{otherwise}
\end{cases} \]

- The Box–Cox encompasses the power, square, and log transformations, depending on the value of \( \lambda \)
- flowClust implements the Box–Cox as part of the model fitting procedure.
Automated Gating With flowClust I


A robust, flexible, model–based approach to automated gating of flow cytometry data.

- Mixture model framework.
- Multivariate-t - robust
- Box–Cox transformation - allows for asymmetric populations.
Automated Gating With flowClust II

Normal-Gamma compound parameterization of the multivariate-$t$

The flowClust Model (Lo et. al.) Complete data log-likelihood

$$\mathcal{L}_c(\Psi \mid y, z, u) = \prod_{i=1}^n \sum_{g=1}^G z_{ig} \log \left\{ \pi_g \phi_p(y_i^{(\lambda)} \mid \mu_g, \Sigma_g / u_i) \right\} \cdot \left| J_p(y_i ; \lambda) \right| \cdot \text{Ga}(u_i, \frac{\nu}{2}, \frac{\nu}{2})$$

- $\Psi = \{ \mu_g, \Sigma_g, \pi_g, \nu, \lambda \}$ population means, covariances, proportions, transformation parameter and degrees of freedom for the $t$–distribution;
- Can be computed efficiently via EM.
flowClust: Model Selection

Standard approach using BIC (Bayesian information Criterion)

$$BIC = -2 \ln(L) + k \ln(n)$$

- fit flowClust models with $G = 1$ through $G = 20$ clusters
- Choose the model with the largest $BIC$ value.
- When $G$ is large, there are many events, or many samples, this becomes time-consuming. Can be parallelized.
Problems with flowClust

G fixed to the "true" number of populations doesn’t necessarily give the best model fit.

Multiple mixture components represent the same cell population.
flowMerge: Modelling Distinct Cell Populations

flowMerge (Finak G, Bashashati A, Brinkman R, Gottardo R. Advances in Bioinformatics, 2009)

Extends the flowClust methodology to identify and model distinct cell populations.

- Merges overlapping mixture components based on entropy.
- Summarizes merged components using a single multivariate-\(t\) distribution based on moment matching conditions.
Mixture Components and Entropy

Entropy measures the uncertainty of a random variable. For mixture models we define the *entropy of clustering* of a G–component mixture model.

**Definition**

Entropy of Clustering

\[
H(G) = -2 \sum_{i=1}^{G} \sum_{j=1}^{N} z_{ij} \log(z_{ij})
\]

- \( z_{ij} \) is the probability that cell \( j \) is assigned to population \( i \).
- Overlapping mixture components: large uncertainty, high entropy.
The flowMerge Algorithm

1. Start with a $\max(\text{BIC})$ flowClust model ($k$ clusters).
2. Compute the entropy for all pairwise model components.
3. Merge the two components that contribute most to the entropy.
4. Recompute the pairwise entropy of the new merged cluster.
5. Repeat from 2. until one component remains.
6. Choose the "best" fitting merged model from the plot of Entropy vs Number of Clusters.
Summarizing Components

We can summarize merged components using the same multivariate–t framework used in flowClust.

\[ p_\ast f_\ast = p_if_i + p_jf_j \]

**Moment Matching Conditions**

\[ p_\ast = p_i + p_j \]

\[ \mu_\ast = \frac{(p_i \mu_i + p_j \mu_j)}{p_\ast} \]

\[ \Sigma_\ast = \frac{(\nu_\ast - 2)p_i}{p_\ast \nu_\ast} \left[ \frac{\nu_i}{\nu_i - 2} \Sigma_i + \mu_i \mu'_i \right] + \frac{(\nu_\ast - 2)p_j}{p_\ast \nu_\ast} \left[ \frac{\nu_j}{\nu_j - 2} \Sigma_j + \mu_j \mu'_j \right] - \frac{(\nu_\ast - 2)p_\ast \mu_\ast \mu'_\ast}{p_\ast \nu_\ast} \]
flowClust vs flowMerge

Figure 4
flowMerge on HSCT and WNV Data

![Graphs showing the flowMerge results on HSCT and WNV Data](image)

- **Entropy of Clustering**
- **Cumulative Number of Merged Observations**
- **WNV Data**
- **FlowMerge: G=5**
- **1.8% misclassified**
flowMerge Caveats

Things to watch out for

- Automated model selection based on a heuristic. Not always a good choice.
- Starting with the max BIC model is sometimes flawed.
- Models with $K = "\text{true \# of clusters}"$ don’t always fit as well as $K = "\text{true number + some outlier clusters}"$.
- Multi-stage gating (gate scatter, then gate fluorescence) may be an invalid assumption.
- Some data sets are just difficult to gate (i.e. WNV)
- Sometimes lacks modelling flexibility (common $\nu, \lambda$).
The Challenges: A reminder

- **Challenge 1**: Fully Automated Algorithms: we know nothing
- **Challenge 2**: Tuned Algorithms: Take a better guess at the number of populations
- **Challenge 3**: The number of populations is predefined.
- **Challenge 4**: The assignment of events to populations is known for some samples.
Gating Strategies Change Depending on the Data

In the flowCAP 2010 algorithm settings and gating strategy, gating strategies change depending on the data.

**Challenge 1**
Automated parameter estimation.
Automated model selection (number of clusters). Two-stage gating: big mistake..

**Challenge 2**
Automated or fixed parameter estimation.
Automated or fixed model selection. Model class dependent on data set.
Gating Strategies Change Depending on the Data II

![Diagram of fluorescence gating]

**Challenge 3**
Gating only fluorescence channels. Automated parameter estimation but fixed number of clusters to known values. (could do better).

**Challenge 4**
Gating only fluorescence channels. Fully automated parameter estimation and model selection.
Overall Impression

Don’t make assumptions
If you know nothing about the data, don’t make too many assumptions. (i.e. Challenge 1: two stage gating - mistake).

One size does not fit all
flowClust is very flexible. More model classes should be explored by default.

Use prior information, but with caution
Models with exactly the "true" number of clusters don’t necessarily provide best fit. Knowing some gating assignments and informative dimensions is better.
Future Improvements

Model Selection

- There is a clear need to improve our model-selection strategy to improve speed and fitting to diverse data.

flowClust with Bayesian Priors

- Include prior information about population locations for gating rare populations or repetitive gating of similar samples.
Acknowledgements

flowCAP Organizing Committee
Ryan Brinkman, BC Cancer Agency
Raphael Gottardo, Fred Hutchinson Cancer Research Center
Richard Scheuermannmm, University of Texas Southwestern Medical Center
Jill Schoenfeld, Tree Star Inc.

Data
BC Cancer Center
Amgen, Inc.
McMaster University
Tree Star Inc.

Funding
NIH