FLAMEMixtureModel Documentation

Module Name: FLAMEMixtureModel
Description: Clusters and estimates population parameters, using preprocessed flow cytometric data
Author: Xinli Hu (Broad Institute), gp-help@broad.mit.edu

Summary:

FLAME (FLow analysis with Automated Multivariate Estimation) uses finite mixture model clustering techniques with novel algorithms and models to identify and characterize discrete populations in flow cytometric data [1]. A pipeline of GenePattern modules implements the method: FLAMEPreviewTransformation, FLAMEPreprocess, FLAMEMixtureModel, FLAMEChooseOptimalClusterNumber, FLAMEMetacluster, FLAMEContourDataGenerator, and FLAMEViewer.

The FLAMEMixtureModel module clusters each preprocessed sample data file over a range of possible cluster numbers. Unlike traditional flow analysis, which relies on sequential 2D gating, FLAMEMixtureModel clusters the data into subpopulations using all dimensions (parameters) at the same time. The user chooses which dimensions to include in the analysis (*channels to cluster* parameter), the density distribution to use (*density* parameter), and a range of cluster numbers (*g min* and *g max* parameters). The module iterates over the range of cluster numbers, clustering the sample data into the specified number of cell subpopulations. It outputs a zip file of all clustering results, which is used as the input file for the FLAMEChooseOptimalClusterNumber module.

Density Distribution:

FLAMEMixtureModel offers four density distributions: normal, t, skew-normal, and skew-t. Normal (Gaussian) distribution is the traditional distribution used in many biomedical data analyses. It assumes the population follows a symmetric distribution that is relatively “dense” or tightly bound. If the true population is more diffuse or contains many outliers, these outer events may be spuriously split off from the population.

The t (Student-*t*) distribution is also symmetric, however allows populations to have heavier “tails”, therefore it is more tolerant to including outlier cells as part of the population.

The skew- normal and skew-t distributions extend the original symmetric distributions by allowing the populations to be skewed in any dimension. These distributions can describe asymmetric populations more accurately.

Computation Time:

The skew distributions require the estimation of more parameters and, therefore, require more computation time. If the populations are expected to be roughly symmetric, t or normal distributions are recommended for faster response.

Computation time also increases with respect to the number of clusters to fit for each sample. If many samples (more than ten) are to be analyzed, first analyze a small number of samples to determine the optimal density distribution and number of clusters. (Choose optimal density distribution and cluster numbers either by reviewing the FLAMEMixtureModel results manually or by analyzing them with the FLAMEChooseOptimalClusterNumber module.) Then, analyze all samples using that density distribution and cluster number (or small range of cluster numbers).

Platform Differences:

The programming libraries used by FLAMEMixtureModel produce slightly different results on Windows and Unix. The generated clusters are the same regardless of platform, but values generated by FLAMEMixtureModel vary slightly across platforms and the number and color associated with each cluster may differ across platforms.

References:

1. Saumyadipta Pyne, Xinli Hu, Kui Wang, Elizabeth Rossin,Tsung-I Lin, Lisa M. Maier, Clare Baecher-Allan, Geoffrey J. McLachlan, Pablo Tamayo, David A. Hafler, Philip L. De Jager, and Jill P. Mesirov. (2009). Automated High-dimensional Flow Cytometric Data Analysis. *PNAS* 106:8519-8524.

Parameters:

Name	Description
preprocessed data	A .zip file containing preprocessed flow sample files in .txt format. This is the output of FLAMEPreprocess module. If FLAMEPreprocess was not used to produce this file, each sample should be in tab-delimited .txt format; where each row is one cell/event, and each column is one antibody/channel. There should be header row denoting the antibody/channel names.
g min	An integer; minimal number of clusters to be fit for each sample. Default is 3.
g max	An integer; maximal number of clusters to be fit for each sample. FLAMEMixtureModel fits a range of cluster numbers between *g min* and *g max*; therefore, *g max* should be greater or equal to *g min*. Default is 10.

density	Density distribution to be used for clustering: normal, t, skew normal, skew t (default).
channels to cluster	A comma-separated list of either channel numbers (e.g., 1, 2, 3, 7) or channel names (e.g., CD4, CD45RA, ZAP70); these denote which antibodies in the (preprocessed) data sample to be used for clustering.
estimate mode	Used only for skew distributions. Whether to estimate the mode for each cluster. Must be the same value used in FLAMEMixtureModel. Default is no. We recommend leaving mode estimation turned off. Using mode estimation for skew distributions is computationally intensive and time consuming. On the GenePattern public server, due to the heavy demand for high-memory servers, such an analysis can potentially take days to complete.
estimation increment	Used only for skew distributions and when estimate mode is set to yes. Step is used to calculate the number of iterations of the Expectation-Maximization(EM) algorithm to perform when computing the Maximum Likelihood(ML) estimate. ML estimation is used to obtain the optimal parameters of the parametric mixture model distribution for each cluster. The value must be greater than 0; the default value is 1. The smaller the increment, the more accurate the estimation, but the slower the estimation step.
seed	An integer; the seed fixes random computational processes during computation. Using the same seed across different runs ensures the reproducibility of results. This is set to a default value of 123456 and does not need to be changed under most circumstances.
output prefix	A prefix for output files.

Output File:

A zip file containing clustering results for each sample at each cluster number. Each set of clustering results is represented in several ways:

1. *.membership.txt: The cluster membership assignment for each cell is recorded as an additional column next to the original data.
2. *.pairplots.png: A panel of 2D plots (including all possible 2D combinations of all antibodies) are plotted as a “paired plot”, where each cluster is denoted by

a distinct color. The values on the x- and y-axis indicate relative intensity levels.

3. pairplots_legend.png: A color key for the clusters in the pair plots.

4. *.parameters.txt: A table of parameters describing all populations within the sample. Each population is quantitatively described by a series of parameters (features) including mean/mode, variance/covariance, degrees of freedom, and skewness. These parameters can be used to quantitatively compare populations across sample groups such as in case/control studies. List of parameters:

    - props: the proportion (a fraction of 1) of cells in the sample belonging to this cluster
    - mus: the location estimation of this cluster; this is the same as the mode only if a symmetric density distribution is used (normal or t).
    - df: degree of freedom
    - Var: a covariate matrix describing how each pair of variables/dimensions change with respect to each other.
    - alpha: skew; a cluster completely symmetric in a certain dimension has an alpha of 0 in the dimension
    - mod: the estimated mode of a cluster

5. *.heatmap.png: A heatmap of mean (or mode, if populations are skew) fluorescent intensities, where each row is one population/cluster and each column is one antibody. Within each column (antibody) red denotes high intensity and blue denotes low intensity.

6. *.locations.txt: A table of mean/mode fluorescent intensities. The last column, weight, gives the percentage of cells in each population.

7. *.ret: A raw output file containing (not in this order):
    a. Mixture model quality assessment criteria, including error code (“0” if no error), Log-likelihood (loglik), AIC, BIC, SWR, and UWR. The last two values are toward the end of the *.ret file.
    b. Cluster parameters, including cluster frequencies/proportions (pro), locations (mu), covariance (sigma), degrees of freedom (dof), skew (delta), and modes (mod). The last value is near the end of the file.
    c. Cluster membership assignments of each cell/event (clust).
    d. Post-posterior probabilities (tao).
    The *.ret file can be accessed in R using the dget() function, or could be read as a text file.

Example Data:

The example data is a subset of the data described in Pyne et al. (2009) [1]:

GenePattern

Input parameter	Value
preprocessed data	Output from FLAMEPreprocess: ftp://ftp.broad.mit.edu/pub/genepattern/example_files/FLAME/SMALLphospho.PreprocessedData.zip
g min	4
g max	6
density	t
channels to cluster	1,2,3,4
estimate mode	yes
estimation increment	0.1

Platform dependencies:

Task type: FLAME
CPU type: Any
OS: Any
Java JVM level:
Language: R (2.7.0 or later)