FLAMEChooseOptimalClusterNumber Documentation

Module Name: FLAMEChooseOptimalClusterNumber
Description: Determines the optimal number of clusters for each sample, using user-specified criterion.
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 define and characterize discrete populations in flow cytometric data [1]. A pipeline of GenePattern modules implements the method: FLAMEPreprocess, FLAMEMixtureModel FLAMEChooseOptimalClusterNumber, FLAMEMetacluster, FLAMEContourDataGenerator, and FLAMEViewer.

The FLAMEChooseOptimalClusterNumber module takes the output of FLAMEMixtureModel module, where each data sample has been clustered over a range of possible cluster numbers, and determines the optimal cluster number (from that range) for each sample. After determining the optimal cluster numbers, the module collects the corresponding result files (membership, parameters, locations, heatmap, pairplots and raw .ret results) for each sample at its optimal cluster number. It outputs a zip file of the collected clustering results, which is used as the input file for the FLAMEMetacluster module.

Assessment Method:

The “optimal” cluster number for biological data can be difficult to assess and in many situation depends on the experimental question. There may be several “optimal” cluster numbers or none. The FLAMEChooseOptimalClusterNumber module offers several methods for determining the optimal cluster number for a sample:

- **Scale-free Weighted Ratio (SWR):** The ratio of average intra-cluster to average inter-cluster scale-free Mahalanobis distances; the averages are weighted to reduce contribution of cluster outliers.
- **Intercluster distance:** The average distance of all pairs of clusters. This measures separation among clusters, and should be maximized for optimal cluster number.
- **Distance ratio:** The ratio of intra-cluster distance to inter-cluster distance. This measures inter-cluster separation as well as within-cluster tightness, and should be minimized for optimal number of clusters.
- **AIC:** Akaike Information Criterion.
- **BIC:** Bayesian Information Criterion.

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**
mixture model	A .zip file containing mixture modeling results of each sample, across a range of cluster numbers. This is the output of FLAMEMixtureModel module.
method	Criterion used for determining optimal number of clusters. Scale-free Weighted Ratio (SWR) (default), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), distance ratio, intercluster distance. Descriptions provided above.
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

- The clustering results for each sample at its optimal cluster number. (The results in this zip file are a subset of the results in the input zip file. For descriptions of the files, see the FLAMEMixtureModel documentation.)
- A text file, *.[mvn, mvt, msn, mst].[assessment method].txt, that lists all samples, the assessment values for all clustering results, and the optimal cluster number chosen for each sample.
- A text file, *.ret, containing the parameters used for the analysis.

GenePattern

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

Input parameter	Value
mixture model	Output from FLAMEMixtureModel: ftp://ftp.broad.mit.edu/pub/genepattern/example_files/FLAME/SMALLphospho.PreprocessedData.MixtureModel.zip
method	Scale-free Weighted Ratio (SWR)
output prefix	SMALLphospho

Platform dependencies:

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