GenePattern

KMeansClustering Documentation

Module name:	KMeansClustering
Description:	K-Means Clustering algorithm
Author:	Marc-Danie Nazaire (Broad Institute), gp-help@broad.mit.edu
Date:	11/15/06
Release:	1.0

Summary:

K-Means clustering is a clustering algorithm that classifies or groups objects into a specified number of clusters. Initially, k cluster centroids (centers) are randomly selected from the given data set and each data point is assigned to the cluster of the nearest cluster center. Each cluster center is then recalculated to be the mean value of its members and all data points are re-assigned to the cluster with the closest centroid. This process is repeated until the distance between consecutive cluster centers converges. The KMeansClustering module can be used to cluster genes(rows) or samples(columns).

References:

- J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

Parameters:

Name	**Description**
input.filename:	Dataset (res, gct, or odf dataset)
output.base.name:	The base name of the output files (.gct)
number.of.clusters:	Number of clusters to use
seed.value:	Seed for the random number generator Specifying the seed value allows you to recreate your results; allowing the program to generate random seed values results in different outcomes even when all other parameters are identical.
cluster.by:	Whether to cluster the dataset by rows or columns
distance.metric:	Algorithm to use to calculate the distance between data points (Note: currently Euclidean is only available)

Return Value:

1. K-Means clustering results in .gct format
2. Stdout.txt: the ”stdout” text output from running the program

Platform dependencies:

Task type:	Clustering
CPU type:	any
OS:	any
Java JVM level:	1.4
Language:	Java