GenePattern

SparseHierarchicalClustering Documentation

Description: Agglomerative hierarchical clustering of genes/experiments, using the sparse hierarchical clustering method of Witten and Tibshirani (2009).

Author: Daniela Witten (Stanford University), dwitten@u.washington.edu

Summary

Standard hierarchical clustering clusters observations using all of the genes. Sparse hierarchical clustering will instead adaptively choose a subset of the genes to use in the clustering. The goal is to (a) identify a small set of genes that are relevant to the clustering, and (b) identify a tighter and less noisy clustering of the observations using only the relevant genes. Each gene will be given a non-negative weight, and depending on the tuning parameter used, many of the genes will have zero weights. If a gene’s weight is zero, then it is not involved in the clustering. The weights of the genes can be used to rank the genes in terms of importance to the clustering (the larger the weight, the more important the gene).

References

Witten DM, Tibshirani R. A framework for feature selection in clustering. *J Am Stat Assoc*. 2010;105:713-726.

Parameters

Name	Description
input.filename (required)	A GCT file containing data to cluster.
method (required)	The type of linking method desired; options are *single*, *complete*, *average*, or *centroid*. Default: *average*.

wbound (required)	The sum of gene weights; this is the tuning parameter that controls the number of genes used in the sparse clustering. Each gene will be given a weight; this tuning parameter is the sum of the weights for the genes. It should be a positive number. The smaller the tuning parameter, the fewer genes have non-zero weights and so the fewer genes involved in the clustering. If it is -1, then the program will automatically choose an optimal tuning parameter. The program runs much more quickly if wbound is specified rather than being set to -1. If set to -1 the suggested value will be calculated and output in the wbound.txt file. Future runs of the same data can use this value to increase the efficiency of the module and decrease runtime. Default: -1.
maxnumgenes (required)	If the number of genes in the data set is very large, then the program can run quite slowly. If a positive integer k is given for maxnumgenes, then only the k genes with highest variance will be used in the analysis. If set to -1, then all genes are used. We recommend setting maxnumgenes to 5000 to speed the analysis on larger data sets. Default: 5000.
cluster.features (required)	If a clustering of the genes with non-zero weights is desired, set to *true* and a GTR file will be output. Default: *false*.
method.features (required)	The type of linkage used to cluster the features, if cluster features is set to *true*. Options are *single*, *complete*, *average*, or *centroid*. Default: *average*.
standardize.arrays (required)	If the arrays should be standardized to have a mean of zero and standard deviation of one before clustering is performed, set to *true*. Default: *true*.

Output Files

1. CDT file

   Contains only the genes that were assigned non-zero weights. Can be viewed (together with the ATR and/or GTR files) using HierarchicalClusteringViewer or JavaTreeview.

2. ATR file

   Contains the clustering of the arrays/observations.

3. GTR file (optional)

   Contains the clustering of the genes. This will be created only if the *cluster.features* option is set to *true*.

4. WEIGHTS.TXT file

   Contains the weight assigned to each gene.

5. WBOUND.TXT file
   Contains some information about the sparse clustering: the value of the tuning parameter wbound used, etc.

Platform Dependencies

Module type:	Clustering
CPU type:	any
OS:	any
Language:	R (2.5 or higher)