



HierarchicalClustering Documentation

Module name: HierarchicalClustering
Description: Agglomerative hierarchical clustering of genes/experiments
Author: Joshua Gould (Broad Institute), gp-help@broad.mit.edu
Date: 7/12/05

HierarchicalClustering is distributed under the license available at <http://rana.lbl.gov/EisenSoftwareSource.htm>.

Summary:

Given a set of items to be clustered (items can be either genes or chips/experiments), agglomerative hierarchical clustering (HC) recursively merges items with other items, or with the result of previous merges, according to their pair-wise distance (with the closest item pairs being merged first). As a result, it produces a tree structure, referred to as dendrogram, whose nodes correspond to: i) the original items (these are the leaves of the tree); and ii) the merging of other nodes (these are the internal nodes of the tree).

HierarchicalClustering will produce a cdt file which contains the original data, but reordered to reflect the clustering. Additionally, either a dendrogram or two dendrogram files are created (one for clustering rows and one for clustering columns). The row dendrogram has the extension gtr, while the column dendrogram has the extension atr. These files describe the order in which nodes were joined during the clustering. For a more detailed description of the format of the output files see <http://genome-www5.stanford.edu/help/formats.shtml>.

The module includes several preprocessing options. The order of the preprocessing operations are:

1. Log Base 2 Transform
2. Row (gene) center
3. Row (gene) normalize
4. Column (sample) center
5. Column (sample) normalize

References:

- M.B. Eisen, et al. "Cluster Analysis and Display of Genome-Wide Expression Patterns," PNAS, 14863-14868 (1998).
- M. J. L. de Hoon, S. Imoto, J. Nolan, and S. Miyano: Open Source Clustering Software. *Bioinformatics*, **20** (9): 1453--1454 (2004).

Parameters:

Name	Description
input.filename	input data file name - .gct, .res, .odf type = Dataset
column.distance.measure	distance measure for column (sample) clustering
row.distance.measure	distance measure for row (gene) clustering
clustering.method	hierarchical clustering method to use
log.transform	log-transform the data before clustering
row.center	whether to center each row (gene) in the data
row.normalize	whether to normalize each row (gene) in the data

GenePattern

column.center
column.normalize
output.base.name

whether to center each column (sample) in the data
whether to normalize each column (sample) in the data
base name for output files

Return Value:

1. cdt file
2. atr file if clustering by columns, gtr file if clustering by rows

Platform dependencies:

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