GenePattern

HierarchicalClustering Documentation

Description: Agglomerative hierarchical clustering of genes/experiments.

Author: Joshua Gould (Broad Institute), gp-help@broadinstitute.org

Summary

Cluster analysis is a means of discovering, within a body of data, groups whose members are similar for some property. Clustering of gene expression data is geared toward finding genes that are expressed or not expressed in similar ways under certain conditions.

Given a set of items to be clustered (items can be either genes or samples), agglomerative hierarchical clustering (HC) recursively merges items with other items or with the result of previous merges, according to the distance between each pair of items, with the closest item pairs being merged first. As a result, it produces a tree structure, referred to as dendogram, whose nodes correspond to:

- the original items (these are the leaves of the tree)
- the merging of other nodes (these are the internal nodes of the tree)

The HierarchicalClustering module produces a CDT file that 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.

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

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

References

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*. 1998;95:14863-14868.

de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. *Bioinformatics*. 2004;20:1453-1454.

Parameters

Name	Description
input filename (required)	Input data file name. File can be in GCT, RES, or ODF formats.
column distance measure (required)	Distance measure for column (sample) clustering. Options include: ● No column clustering ● Uncentered correlation: The same as the Pearson correlation, except that the sample means are set to zero in the expression for uncentered correlation. The uncentered correlation coefficient lies between –1 and +1; hence the distance lies between 0 and 2. ● Pearson correlation (default): Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. It is a measure for how well a straight line can be fitted to a scatter plot of x and y. If all the points in the scatter plot lie on a straight line, the Pearson correlation coefficient is either +1 or -1, depending on whether the slope of line is positive or negative. If it is equal to zero, there is no correlation between x and y. ● Uncentered correlation, absolute value: The same as the absolute Pearson correlation, except that the sample means are set to zero in the expression for uncentered correlation. The uncentered correlation coefficient lies between 0 and +1; hence the distance lies between 0 and 1. ● Pearson correlation, absolute value: The absolute value of the Pearson correlation coefficient is used; hence the corresponding distance lies between 0 and 1, just like the correlation coefficient. ● Spearman's rank correlation: Nonparametric version of the Pearson correlation that measures the strength of association between two ranked variables. To calculate the Spearman rank correlation, each data value is replaced by their rank if the data in each vector is ordered by their value. Then the Pearson correlation between the two rank vectors instead of the data vectors is calculated. It is useful because it is more robust against outliers than the Pearson correlation. ● Kendall's tau: The Kendall tau distance is a metric that counts the number of pairwise disagreements between two lists. The larger the distance, the more dissimilar the

	two lists are. • Euclidean distance: Corresponds to the length of the shortest path between two points. Takes into account the difference between two samples directly, based on the magnitude of changes in the sample levels. This distance type is usually used for data sets that are normalized or without any special distribution problem. • City-block distance: Also known as the Manhattan or taxi cab distance; the city-block distance is the sum of distances along each dimension between two points.
row distance measure (required)	Distance measure for row (gene) clustering. Options include: • No row clustering (default) • Uncentered correlation • Pearson correlation • Uncentered correlation, absolute value • Pearson correlation, absolute value • Spearman's rank correlation • Kendall's tau • Euclidean distance • City-block distance **NOTE: Filtering beforehand is recommended since row clustering is computationally intensive.**
clustering method (required)	Hierarchical clustering method to use. Options include: • Pairwise complete-linkage (default): The distance between two clusters is computed as the maximum distance between a pair of objects, one in one cluster and one in another. • Pairwise single-linkage: The distance between two clusters is computed as the distance between the two closest elements in the two clusters. • Pairwise centroid-linkage: The distance between two clusters is computed as the (squared) Euclidean distance between their centroids or means. • Pairwise average-linkage: The distance between two clusters is computed as the average distance between the elements in the two clusters.
log transform	Specifies whether to log-transform the data before clustering. Default: no

row center	Specifies whether to center each row (gene) in the data. Centering each row subtracts the row-wise mean or median from the values in each row of data, so that the mean or median value of each row is 0. Default: no
row normalize	Specifies whether to normalize each row (gene) in the data. Normalizing each row multiplies all values in each row of data by a scale factor S so that the sum of the squares of the values in each row is 1.0 (a separate S is computed for each row). Default: no
column center	Specifies whether to center each column (sample) in the data. Centering each column subtracts the column-wise mean or median from the values in each column of data, so that the mean or median value of each column is 0. Default: no
column normalize	Specifies whether to normalize each column (sample) in the data. Normalizing each column multiplies all values in each column of data by a scale factor S so that the sum of the squares of the values in each column is 1.0 (a separate S is computed for each column). Default: no
output base name (required)	Base name for the output files.

Output Files

1. CDT file
   Contains the original data, but reordered to reflect the clustering.
2. ATR file (if clustering by columns/samples) or GTR file (if clustering by rows/genes)
   These files describe the order in which nodes were joined during the clustering.

Platform Dependencies

Module type: Clustering

CPU type: any

OS: any

Language: Java 1.5 and higher; C

GenePattern Module Version Notes

Version	Date Released	Description
5	2/10/2009	
6	3/5/2013	Updated for Java 7.