HierarchicalClustering Documentation

Module name: HierarchicalClustering
Description: Agglomerative hierarchical clustering of genes/experiments
Author: Stefano Monti (Broad Institute) smonti@broad.mit.edu
Date: 4/16/2003
Release: 1.0

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 dendogram, 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). If k clusters are required (k 1), the merging proceeds until k nodes are left. Two distance measures can be used: 1) the Euclidean distance; 2) one minus the Pearson correlation. The output produced consists of at most two output files (depending on the format of the argument to the option '-o'. See below):
1. A '.clu' file in the format (assuming k clusters are produced):
    1:
    $item_1$ $item_2$ … $item_{n1}$
    2:
    $item_1$ $item_2$ … $item_{n2}$
    ..
    k:
    $item_1$ $item_2$ … $item_{nk}$
2. A '.gwt' file with the description of the sub-tree for each cluster. The description of a cluster has the following regular expression format:
    Cluster → Node
    Node → [name (Node, Node {distance})]
    Node → item
    Name → alphanumeric string

References:

- M.B. Eisen, et al. "Cluster Analysis and Display of Genome-Wide Expression Patterns," PNAS, 14863-14868 (1998).

Usage/Example:
```
HC.out <- HierarchicalClustering("ALB_ALT_AML.gct", data.format=0, normalize.type=1, num.iter=0, output.name="HCout.clu", num.classes=1, merge.type="average")
```

Parameters:

Name	Description	Choices
input.filename	The data based on which to carry out the clustering	It can be a '.gct', or a '.res', or a tab-delimited text file with or without row and column headers
data.format	The format of input.filename	0='.gct' or '.res' format (default); 1=row/column headers; 2=raw data (no headers)

normalize.type	Type of normalization to perform on data	1=row normalize (default); 2=column normalize; 3=both
num.iter	Number of row/column normalization to perform. It supercedes normalize.type.	A non-negative integer
ouput.name	Where to save the output	Can be any alphanumeric string. If it ends in '.clu', only the '.clu' output is saved (see Summary above). If it ends in '.gwt', only the '.gwt' output is saved. Otherwise, output is saved in the two files <output.name>.clu and <output.name>.gwt.
num.classes	Number of clusters to return	An integer k, such that 1 k N, where N is the number of items to be clustered.
merge.type	How to update the distance measure following the merging of two nodes.	"average", "single", "complete".

Return Value: An R list with components:

1. <output.name>.clu see description in Summary.
2. <output.name>.gwt see description in Summary.

Platform dependencies:

Task type:	Clustering
CPU type:	any
OS:	any
Java JVM level:	1.4
Language:	Java
Support files:	none

Native command line: <java> <java_flags> -cp <libdir>file_support.jar<path.separator><libdir>ui_support.jar<path.separator><libdir>gp-common.jar<path.separator><libdir>hcl.jar edu.mit.wi.genome.geneweaver.clustering.HierarchicalClustering <input.filename> -h<data.format> -n<normalize.type> -N <num.iter> -o <output.name> -K <num.classes> -L <merge.type> -r -s -c<cluster.by>