GeneNeighbors Documentation

Module name: GeneNeighbors
Description: Select features that most closely resemble or differ from another feature.
Author: Ken Ross (Broad Institute) gp-help@broadinstitute.org
Date: 10/23/03
Release: 1.0

The GeneNeighbors module calculates the nearest neighbors for a particular gene (or other continuous vector pseudo-gene) by trying to find other genes whose expression values follow similar trends for the samples. The user specifies the number of nearest neighbors to find for a particular gene by entering a value for the num neighbors parameter. There are four choices for the distance metric: Cosine, Euclidean, Manhattan, and Pearson. The cosine distance is given by

$$d_C = \sum_i x_i * y_i \left/ \left(\sum_i x_i^2 * \sum_i y_i^2 \right)^{1/2} \right.$$ where i is the sample number, x_i is the named

reference gene's expression value for sample i, and y_i is the expression value of the

gene we are testing. The Euclidean distance is given by $d_E = \left(\sum_i (x_i - y_i)^2 \right)^{1/2}$ where i

is the sample number, x_i is the named reference gene's expression value, and y_i is the expression value of the gene we are testing. The Manhattan distance is given by $d_M = \sum_i |x_i - y_i|$ where i is the sample number, x_i is the named reference gene's

expression value, and y_i is the expression value of the gene we are testing. The Pearson distance is calculated by

$$d_P = n\left(\sum_i x_i * y_i \right) - \left(\sum_i x_i \right) * \left(\sum_i y_i \right) \left/ \left(\left[n\sum_i x_i^2 - \left(\sum_i x_i \right)^2 \right] * \left[n\sum_i y_i^2 - \left(\sum_i y_i \right)^2 \right] \right)^{1/2} \right.$$

where i is the sample number, n is the number of samples, x_i is the named reference gene's expression value, and y_i is the expression value of the gene we are testing. Running this algorithm produces a table with two columns: 1) Feature - contains the gene's identifier from the input file, 2) Score - contains the calculated distance for the gene relative to the reference gene. The genes in the output table are ordered based upon the score.

The results table from the GeneNeighbors algorithm can be viewed with the GeneListSignificanceViewer and the data results file can be viewed with the HeatMapViewer.

References:

- Golub T.R., Slonim D.K., et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, 531-537 (1999). and the supplemental information on the website http://www-

genome.wi.mit.edu/cgi-bin/cancer/publications/pub_menu.cgi for a more complete description of marker permutation testing.

- Slonim, D.K., Tamayo, P., Mesirov, J.P., Golub, T.R., Lander, E.S. (2000) Class prediction and discovery using gene expression data. In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB) 2000. ACM Press, New York, pp. 263–272.

Parameters:

Name	Description	Choices
Dataset	input dataset - .gct, .res, odf type=Dataset	
gene.accession	reference gene accession from dataset to find neighbors for	
num.neighbors	number of neighbors to find. Default value: 50	
marker.list.file:	output filename for analysis results - .odf file	
marker.dataset.file	output filename raw data for selected markers (.gct)	
distance.metric	metric for finding neighbors. Default value: Pearson	Cosine Distance, Euclidean Distance, Manhattan Distance, Pearson Correlation
filter.data	if no, values below will be ignored. Default value: no	yes/no
min.threshold	minimum threshold for data. Default value: 20	
max.threshold	maximum threshold for data. Default value: 20000	
min.fold.diff	minimum fold difference for filtering genes. Default value: 5	
min.abs.diff	Minimum absolute difference for filtering genes. Default value: 50	

Output Files:

1. marker.list.file: output file (.odf format) with table of analysis results.
2. marker.dataset.file: output file (gct format) with raw data for top markers.

Platform dependencies:

Module type: GeneListSelection
OS: any
Java JVM level: 1.5
Language: Java