ClassNeighbors Documentation

Module name: ClassNeighbors
Description: Select genes that most closely resemble a profile
Author: Ken Ross (Broad Institute) kross@broad.mit.edu
Date: 10/23/03
Release: 1.0

Summary: The ClassNeighbors tool performs marker analysis that helps the user determine which genes are most closely correlated with a class template and how significant that correlation is for each gene. The user chooses the dataset and class template, and also the number of genes they want to consider for correlation. Genes are ranked by the user selected method of either the Signal-to-Noise-Ratio (SNR) or the t-Test. The Signal-to-Noise feature selection method looks at the difference of the means in each of the classes scaled by the sum of the standard deviations: $Sx = (\mu_0 - \mu_1)/(\sigma_0 + \sigma_1)$ where μ_0 is the mean of class 0 and σ_0 is the standard deviation of class 0. The t-Test statistic is the same as the Signal-to-Noise except the denominator is $(\sigma_0^2 + \sigma_1^2)^{1/2}$, this would be the statistic used in the T-test. Note that $(\sigma_0 + \sigma_1) > (\sigma_0^2 + \sigma_1^2)^{1/2}$ always; also these two statistics are identical when $\sigma_0 = 0$ or $\sigma_1 = 0$. Thus, the Signal-to-Noise statistic penalizes genes that have higher variance in each class more than those genes that have a high variance in one class and a low variance in another. This bias is perhaps useful for biological samples, e.g. in a case of tumor versus normal where in one class, the gene is working normally and regulated relatively strictly, and in the other class the gene is broken and varying more widely. Running this algorithm causes the number of neighbors most correlated with each class to be calculated for each of the classes and performs a permutation test to assess the significance of the score for each gene. The permutation test can be used to calculate whether the top marker genes with respect to a biologically meaningful phenotype (e.g. morphology) are statistically significant. We do this by comparing the signal-to-noise scores for top marker genes with the corresponding ones for random permutation versions of the class labels (phenotype). This test permutes the class assignments N times (where N corresponds to the value in the Num.permutations entry). For each permutation, the genes are ranked. Then a histogram of signal-to-noise scores for each rank is built. For example, one histogram for all N top markers (k=1), another histogram for all N second best (k=2), etc. These histograms represent a reference statistic for the best match, second etc. and for a given value of k different genes contribute to it. Notice that the correlation structure of the data is preserved by this procedure. Then for each value of k one determines the 1%, 5% and user set (from the User.pval entry) significance levels. This test controls for both the number of genes ranked (the more genes ranked, the greater chance there will be one with a high score to the random template) and also for correlation between genes. The algorithm produces an output table that has 12 columns with the following meanings:

1. This column with heading # contains an index for the rows in the table. It is useful when you want to sort to the original ordering for the marker analysis results.
2. The Class column contains the class label for which the gene in each row is more highly expressed. The class label is either a 0 or 1 or the label specified in the CLS file on the line with the '# class0 class1' information.
3. The Score column contains the absolute value of the signal-to-noise ratio for the row's gene.
4. The Mean0 column specifies the mean of the gene in the class 0 samples.
5. The Std0 column specifies the standard deviation of the class 0 samples.

6. The Mean1 column specifies the mean of the gene in the class 1 samples.
7. The Std1 column specifies the standard deviation of the class 1 samples.
8. The Perm 1 % column contains for each gene the signal-to-noise ratio for the one percent level from the permutation of the class labels.
9. The Perm 5 % column contains for each gene the signal-to-noise ratio for the five percent level from the permutation of the class labels.
10. The Perm (user) column contains for each gene the signal-to-noise ratio for the user set p-value (default 0.5) from the permutation of the class labels.
11. The Feature column gives the name for the row's gene where the name comes from the input data file.
12. The Desc column gives the description (if any) of the gene where the description comes from the input data file.

Within the table, the rows are sorted by the class that the markers are correlated with followed by the values of the signal-to-noise ratio.

Our implementation of the ClassNeighbors algorithm also includes several basic data pre-processing options. The thresholding option allows the user to set minimum and maximum thresholds for the data. Any value in the data set that is less then the value for the minimum threshold is set to the minimum threshold value. Similarly, any value that is greater than the maximum threshold is set to the maximum threshold value. There is also a variation filter that will remove rows from the data set whose values do not vary greatly. For a given row (gene), minVal is the minimum value in that row and maxVal is the maximum value in that row. If maxVal/minVal is greater than the specified minimum fold difference variation ratio (defaults to 5) and maxVal - minVal is greater than the specified minimum absolute difference (defaults to 50), then the row passes the filter. Any rows that do not pass the filter are excluded from the list of features that can be used for the ClassNeighbors gene ranking algorithm. If more neighbors are requested than two times the number of genes remaining after filtering, an error will be produced.

The results table from the ClassNeighbors 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.

Usage/Example:

```
function(data.filename, class.filename, marker.gene.list.file,
       marker.data.set.file, num.neighbors="50",
       num.permutations="100", user.pval="0.5", min.threshold="10",
       max.threshold="16000", min.fold.diff="5", min.abs.diff="50",
       mean.or.median="-d", ttest.or.snr="-S", server=defaultServer)
{
   return (runAnalysis("ClassNeighbors", data.filename=data.filename,
         class.filename=class.filename,
```

```
        marker.gene.list.file=marker.gene.list.file,
        marker.data.set.file=marker.data.set.file,
        num.neighbors=num.neighbors,
        num.permutations=num.permutations, user.pval=user.pval,
        min.threshold=min.threshold, max.threshold=max.threshold,
        min.fold.diff=min.fold.diff, min.abs.diff=min.abs.diff,
        mean.or.median=mean.or.median, ttest.or.snr=ttest.or.snr,
        server=server))
}
```

Parameters:

Name	Description
data.filename:	data file (.res, .gct or .odf)
class.filename:	class file (.cls, .odf)
marker.gene.list.file:	output filename for analysis results (.odf)
marker.data.set.file:	output filename for marker data (.gct)
num.neighbors:	number of neighbors to find
num.permutations:	number of permutations in permutation test
user.pval:	user-set significance value for permutation test
min.threshold:	minimum threshold for data
max.threshold:	maximum threshold for data
min.fold.diff:	minimum fold difference for filtering genes
min.abs.diff:	minimum absolute difference for filtering genes
mean.or.median	use mean or median values for feature selection
ttest.or.snr	use the t-test or signal to noise ratio

Return Value: An R list with components:

1. marker.gene.list.file: output file with table of analysis results.
2. marker.data.set.file: output file (gct format) with raw data for selected markers.
3. Stdout: the ”stdout” text output from running the program.
4. Stderr: the “stderr” error report from running the program.

Platform dependencies:

Task type:	GeneListSelection
CPU type:	any
OS:	any
Java JVM level:	1.3
Language:	Java
Support files:	none

Native command line: <java> <java_flags> -cp <libdir>trove.jar<path.separator><libdir>Jama-1.0.1.jar<path.separator><libdir>broad-cg.jar<path.separator><libdir>jakarta-oro-2.0.8.jar<path.separator><libdir>RunMarkerSelection.jar edu.mit.wi.genome.expresso.alg.RunMarkerSelection.RunMarkerSelection -t <data.filename> -c <class.filename> -H -m <min.threshold> -M <max.threshold> -f -F <min.fold.diff> -D <min.abs.diff> -o <marker.gene.list.file> -p -P <num.permutations> -N <num.neighbors> -l <user.pval> -O <marker.data.set.file> -G <filter.data> <mean.or.median> <ttest.or.snr>