GenePattern

ComparativeMarkerSelection Documentation

Module name:	ComparativeMarkerSelection
Description:	Compare different approaches to marker selection
Author:	Joshua Gould, Gad Getz, Stefano Monti (Broad Institute) gp-help@broad.mit.edu
Date:	January 31, 2005
Release:	1.0

The ComparativeMarkerSelection module includes several approaches to determine the features that are most closely correlated with a class template and the significance of that correlation. The module outputs a file containing the following columns:

1. ***Rank*** - The rank of the feature within the dataset based on the value of the test statistic. If a two-sided p value is computed, the rank is with respect to the absolute value of the statistic.
2. ***Feature*** - The feature name.
3. ***Score*** - The value of the test statistic.
4. ***Feature P*** - The feature-specific p value based on permutation testing.
5. ***FPR*** (False Positive Rate) – The expected proportion of null hypotheses/features having a score better than or equal to the observed one. This measure is not feature-specific, since it is computed by counting the proportion of features having a permuted score better than or equal to the given feature’s observed score.
6. ***FWER*** (Family Wise Error Rate) - the probability of at least one null hypothesis/feature having a score better than or equal to the observed one. This measure is not feature-specific (see comment about FPR in [5]).
7. ***Rank P*** - The rank-specific p value measures the probability that a null hypothesis/ feature with the given *rank* (i.e., a hypothesis/feature whose observed score has that rank) will be rejected. This measure is not feature-specific since it is computed by comparing the permuted scores of equally ranked features, which can be different at different permutation iterations.
8. ***FDR (BH)*** - An estimate of the false discovery rate by the Benjamini and Hochberg procedure (3). The FDR is the probability of a rejected hypothesis being null.
9. ***Bonferroni*** - The value of the Bonferroni correction applied to the feature specific p value.
10. ***Q Value*** - An estimate of the FDR using the procedure developed by Storey and Tibshirani (4). See the definition of the FDR in [8].

The results from the ComparativeMarkerSelection algorithm can be viewed with the ComparativeMarkerSelectionViewer.

Parameters:

Name	**Description**	**Choices**
input.filename	The input file - .res, .gct, .odf	
cls.filename	The class file - .cls	
test.direction	The test to perform (up-regulated for class 0, up-regulated for class 1, two-sided)	Class 0;Class 1; 2 Sided
test.statistic	The statistic to use	SNR;T-Test;SNR

		(median);T-Test (median);T-Test (min std)
min.std	The minimum standard deviation if test statistic is T-Test (min std)	
number.of.permutations	The number of permutations to perform	100
complete	Whether to perform all possible permutations	yes;no
balanced	Whether to perform balanced permutations	yes, no
fix.standard.deviation	Whether to adjust the standard deviation, as is done in GeneCluster	yes;no
output.file	The name of the output file	
random.seed	The seed of the random number generator	

Return Value:

1. An odf file containing the results

References:

1. 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.

2. 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.

3. Benjamini, Y., Hochberg, Y. (1995). " Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing ", *Journal of the Royal Statistical Society B,* 57 289-300.

4. Storey JD and Tibshirani R. (2003) Statistical significance for genome-wide experiments. Proceedings of the National Academy of Sciences, 100: 9440-9445.

Platform dependencies:

Task type:	GeneListSelection
CPU type:	any
OS:	any
Java JVM level:	1.4
Language:	Java, R
Support files:	Jama-1.0.1.jar, broad-cg.jar, colt.jar, MarkerSelection.jar

Native command line:

```
<java> -DR\=<R_HOME> -Dlibdir\=<libdir> <java_flags> -cp <libdir>Jama-1.0.1.jar<path.separator><libdir>broad-cg.jar<path.separator><libdir>colt.jar<path.separator><libdir>MarkerSelection.jar edu.mit.broad.marker.MarkerSelection <input.filename> <cls.filename> <number.of.permutations> <test.direction> <output.file> <balanced> <complete> <fix.standard.deviation> <test.statistic> <random.seed> <min.std>
```