BOOTSTRAPPED ROC (BROC)

Statistical analysis of gene expression

PURPOSE

The bROC algorithm (version 3) is used in the discovery of differentially expressed probes/genes in microarray and RNA-seq experiments.

bROC plugin deploys in CLC Main Workbench and CLC Genomics Workbench.

BENEFITS

- Works particularly well for experiments with a small number of experimental/biological replicates.
- Includes data normalization (for RNA-seq, mainly).
- When combined with *RNA-seq Analysis (CLC bio)*, provides complete differential expression analysis workflow for RNA-seq data.
- Graphical outputs facilitate interpretation of results.
- With no user-selectable parameters, the algorithm is easy to use.
- Non-parametric approach is applicable to all platforms producing expression data for a large number of features (transcripts, genes).

ROC ANALYSIS

ROC (receiver operating characteristic) is a generally applicable, non-parametric procedure that provides insight into the discriminatory properties of data features for a binary classifier. However, the method is not efficient for gene expression experiments as they generally do not produce a sufficient number of samples. bROC overcomes this limitation by resampling (bootstrapping) the expression data to produce a large number of simulated measurements and detailed curves of sensitivity (probability of true positive detection) vs. 1-specificity (probability of false positive detection) for all features of interest. The area under the curve (AUC) is the primary statistic used for identifications of regulated features (genes/probes).

INPUT DATA

The bROC compares expression between two biological states/endpoints (e.g., treatment and control samples, disease and normal, etc.). At least two experimental/biological replicates are required per state. The algorithm is especially useful for analysis of data sets with small number of replicates and large number of features/probes (thousands or tens of thousands).

If needed, the input data are automatically log2 transformed before they are used in ROC analysis. For RNA-seq data, which contain null values, the typically used automatic transforms are unity shift and log2.

OUTPUTS

For each experimental feature, bROC produces the following statistics (discrimination scores):

- *CONF = 2AUC–1,* where *AUC* is area under the curve. *CONF* = 1 (*AUC* = 1) indicates perfect separation of the expression measurements between two states and *CONF* = 0 (*AUC* = 0.5) indicates no separation (no differential expression).

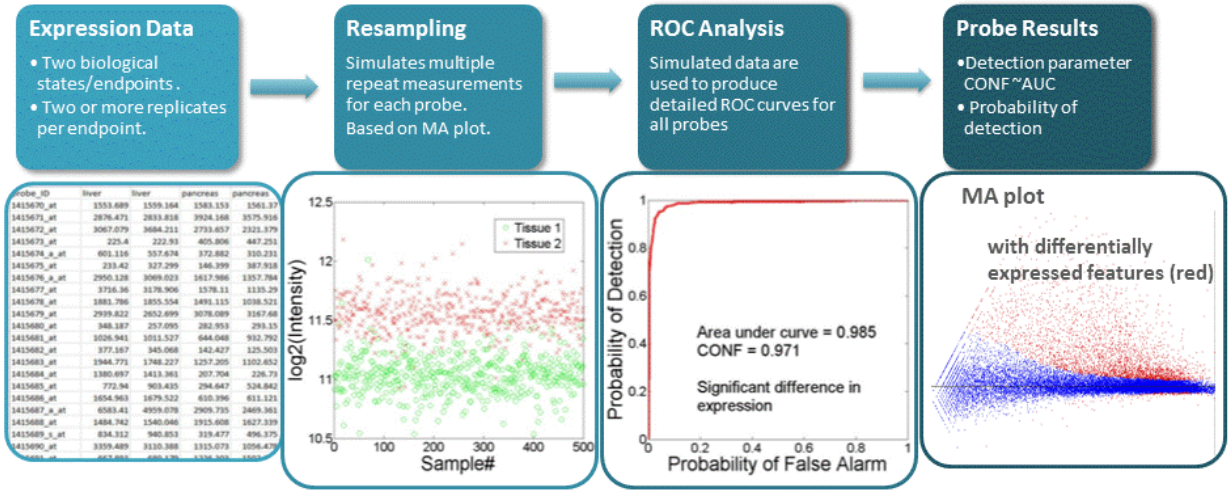
  Typically, features with CONF ≥ 0.95 are considered to be differentially expressed.

- *PD* (probability of detection) balanced against *PFA* (probability of false alarm), with PFA ≅ 1 – PD. This value is calculated at the intersection of ROC curve and diagonal of the ROC plot.

Other outputs include estimates of standard deviation for CONF and PD, and fold change used in the ROC evaluation (for transformed and/or normalized data).

Differentially expressed features are depicted on MA and XY scatter plots. The 'volcano' plot shows CONF vs. Fold Change.

ANALYSIS FLOW



BioFormatix

12396 World Trade Drive, Suite 315, San Diego, CA 92128, USA