LRPath

Pathway Analysis using Logistic Regression

Methods for populating the concepts tested

The method used to determine the concept types/databases and their assigned genes is species-dependent.

Human: All data is from the database of ConceptGen (http://conceptgen.ncibi.org)

Mouse and Rat: For cytoband, data is from the R package *org.Mm.eg.db* for mouse or *org.Rn.eg.db* for rat. For all other concept types, Homologene from NCBI is used to obtain the human homologs and the ConceptGen database (http://conceptgen.ncibi.org) is used.

Drosophila, yeast, c. elegans, and zebrafish: Only KEGG, Gene Ontology, and cytoband are available. The appropriate species specific R package is used, of the form *org.Xx.eg.db*, where *Xx* refers to the species.

LRpath Statistical Method

LRpath functionally relates the odds of gene set membership (dependent variable) with the statistical significance of differential expression (independent variable) using logistic regression, and calculates q-values using the FDR method as a measure of statistical significance. The basic question asked by LRpath is, "Does the odds of a gene belonging to a pre-defined gene set increase as the significance of differential expression increases?" Logistic regression is a natural extension of the Chi-squared test, allowing the significance values to remain on a continuous scale and not requiring the use of significance thresholds. LRpath displays robust behavior and improved statistical power compared to tested alternatives[1].

Detailed Methods are provided here and in the original publication:

Suppose that for a given microarray experiment we have assigned the statistical significance of the comparison of interest to each gene in terms of *p*-values. Our logistic regression method proceeds as follows. For each category (i.e. gene set) *c,* the dependent variable *y* is defined as 1 for genes in *c*, and 0 for all other genes. We use the significance statistics, defined as –log(*p*-values), as the explanatory variable *x*, although a different significance measure could be used. If π is the proportion of genes belonging to the category (*y*=1) at a specified *x* value, then π / (1- π) are the corresponding odds that a gene with significance *x* is a member of this particular category. If the log odds value increases as *x* increases, then we conclude the category is associated with the differential expression. Logistic regression is used to model the log-odds of a gene belonging to the specific category as a linear function

of the statistical significance *x*:

where α is the intercept, β is the slope, and both α and β are estimated from the data. The slope parameter, β, corresponds to the change in the log odds of belonging to the specific category for a unit increase in *x* (or ten-fold decrease in p-value). When $\beta > 0$ we conclude that the category of interest is

$$W = \left(\frac{\hat{\beta}}{} \right)^2$$

"enriched" with differentially expressed genes (or conversely that the category is "depleted" if β <0). The evidence in the data that β > 0 (or < 0) for a specific category is assessed by calculating the *p*-value for the null hypothesis that β = 0 against the alternative that $\beta \neq 0$ based on the maximum likelihood parameter estimates and the Wald test. The Wald statistic, W, is calculated as:

where $\hat{\beta}$ is the maximun likelihood (ML) estimate for β, s_{β} is the standard error of $\hat{\beta}$, and the ML is estimated using the iteratively weighted least squares (IWLS) algorithm. For testing β =0, *W* can be shown to follow a chi-square distribution with one degree of freedom and the *p*-value is calculated assuming this null distribution. The *p*-values from the test of each category *c*, are then adjusted for multiple testing controlling the false discovery rate (FDR) (Benjamini and Hochberg 1995) . Most likely enriched gene sets will be identified based on the p-value, or based on the odds ratio if a ranking independent of category size is desired.

Advanced Analysis Options

Maximum number of genes in concept: The default is to include all concepts in the analysis no matter how large. One may wish to limit the maximum size in order to avoid very large, often vague or broadly defined concepts. An example would be the GO term "cytoplasm". The database in ConceptGen contains concepts that range in size from 5 genes to 1000 genes.

Minimum number of genes in concept: The default is to limit testing to concepts that contain at least 10 genes with a value provided. This is the recommended minimum size for logistic regression. However, users may choose to use a different cutoff.

Low value used to calculate odds ratio: The provided p-value to use for the lower value to calculate the odds ratio of enrichment. The default is 0.001.

High value used to calculate odds ratio: The provided p-value to use for the upper value to calculate the odds ratio of enrichment. The default is 0.5.

Significance cut-off for reporting the driving genes: The p-value used to determine which genes will be reported in the right-most column of the output as differentially expressed. The default is 0.05.

Reference

[1]Sartor MA, Leikauf GD, Medvedovic M. LRpath: A logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics. 2009; 25(2): 211-7.