Gene Expression Analysis with Groups

Video Script

Associated with "EcoCyc-Groups_Gene-Expression-Analysis_081913.mov"

This webinar will discuss the analysis of gene expression datasets with Web Groups.

We assume that you are starting with a gene group containing a set of genes of interest, such as a set of genes that an external statistical analysis program has identified as significantly up- or down-regulated, compared to a control. Note that the group should have been created with respect to the BioCyc organism of interest,, which in our case is E. coli. As our example we will use a set of genes that were the most significantly up- and down-regulated during growth of E. coli on a tryptophan-rich medium compared to a minimal medium lacking tryptophan.

The Groups Transformations webinar already presented one analysis approach, namely to transform a set of genes to the set of pathways that it contains. We present two additional approaches here: visualization of a gene set on an organism's metabolic map diagram, and statistical enrichment analysis of the gene set.
[set of genes to use as an example: http://biocyc.org/group?id=Biocyc11-61-3553016696]

Visualizing Genes on a Cellular Overview

We call an organism's metabolic map diagram in BioCyc the cellular overview. To display a set of genes on the cellular overview, make sure the gene column is selected, go to the groups menu, and click on "paint data on cellular overview." That brings up the cellular overview with the genes that are in our group highlighted (flashlight effect on green highlighting) so now we can visually explore which areas of the cellular overview are highlighted. If we zoom in far enough, it will start to show you the names of some of these pathways. So for instance, here's a pathway that contains a few highlighted genes.

If we mouse over a reaction or metabolite, we'll see a tooltip that identifies that entity. If we click "Keep Open," the tooltip window will stay open, and we can move it around and create additional windows for use in a publication. The "E, R, P" buttons within reaction tooltips let us control how much information is presented.

The control panel at the right lets us remove or add back highlighting for multiple sets of genes that we might have highlighted at one time.

Please also be aware that larger sets of genes, along with quantitative expression values for multiple time points, can be painted onto this diagram, as described in the Omics Data Analysis section of the BioCyc Website User's Guide [show how to navigate to that page on web site].

Enrichment Analysis

The next way we will analyze our same group of genes is through enrichment analysis. Enrichment analysis is a statistical technique for finding what a set of genes or other objects have in common with one another. That is, do some of the genes in our group belong to biologically related sets of genes at statistically significant levels? For example, enrichment analysis could tell you that a set of genes contains more genes involved in cell division than you would expect to find by chance.

Web groups offers a variety of enrichment analysis tools that are available from the enrichment menu. For instance, given a set of genes, we can compute whether that set is enriched for the presence of metabolic pathways, that is, to find pathways that contain more genes from our set than we would expect to occur by chance. To do so, we use the Enrichments menu, and select "Genes enriched for pathways", which pops up a menu with options that control the details of the statistical computation. We'll just take defaults, although multiple options are available for the statistical test to use and for correction of multiple-hypothesis testing.

This analysis will create a new group of pathways. Some things to note about this group: the first column contains both individual pathways and pathway classes, which are distinguished by the use of uppercase initial letters. The second column is the p-value meaning the statistical significance, and the column is ordered by p-values, so the pathways at the top are the ones that are most over represented. The last column shows you the genes from the original set that are present in this pathway. I'm going to delete that match column right now just so we can see the pathways a little better.

We see that these are all the pathways that the enrichment analysis has produced and we can see that a lot of them relate to biosynthesis of amino acids so that's something we've learned.

Next we will perform a related enrichment analysis on the same starting gene set, namely to simultaneously search for enriched sets of Gene Ontology terms, pathways, and transcriptional regulators. To do so, we select a different choice under the Enrichments menu. In the resulting group, the first column contains a mixture of Gene Ontology categories, pathways, and transcriptional regulators. As we scroll down the list we see that our gene list contained many genes involved in the GO category "SOS response," such as recN. Further down we see that our original gene list is statistically enriched for genes that are regulated by the transcriptional regulators lexA and ecpR.