Computer-Aided Discovery Methods

Computational exercise: Transcriptional regulatory pathways

Chad Creighton and Chris Miller

Spring 2009

Purpose

The goal of this exercise is to show how we can relate the results of two independent gene expression datasets to each other. In the case where one datasets provides "transcriptional signatures" of known oncogenic pathways, we can get clues as to which oncogenic pathways may be represented within the results obtained from another dataset.

Goal

Determine what oncogenic pathway gene signatures are represented by a gene expression signature of IGF.

Datasets

Two public gene expression profile datasets have been made available for this exercise:

- The "IGF dataset," which consists of MCF-7 breast cancer cells treated at 3hr and at 24hr with or without the IGF-I molecule.
- The "Bild oncogene dataset" of of breast cells activated with five different oncogenes (Myc, Ras, E2F3, beta-catenin, Src).

Software

This exercise may be carried out using the following software:

- Excel
- R
- GSEA (publicly available from the Broad Institute http://www.broad.mit.edu/gsea/)

For the parts of the exercise using Excel or R, students are welcome to use other software (if they prefer) in order to accomplish the required tasks.

Instructions

1) From the IGF dataset, define a set of genes to represent and "IGF signature." Genes making up the signature are to be the following:
    a. Genes that are HIGHER than control at BOTH the 3 hr and the 24 hr time point, with P<0.01 by two-sided t-test for each time point (for each time point, compare IGF treated with the corresponding control, use the TTEST and AVERAGE functions in Excel).
    b. Genes that are LOWER than control at BOTH the 3 hr and the 24 hr time point, with P<0.01 by two-sided t-test for each time point.
2) From the Bild dataset, define a gene signature (with a set of "up" genes and a set of "down" genes) for each of the five oncogenes represented in the dataset. Use the following statistical cutoffs in defining the signatures:
    a. Myc signature: P<0.001 by two-sided t-test, comparing the Myc group with GFP control
    b. Ras signature: P<0.00001 by two-sided t-test
    c. E2F3 signature: P<0.001 by two-sided t-test
    d. beta-catenin signature: P<0.001 by two-sided t-test
    e. Src signature: P<0.0001 by two-sided t-test
    (Don't worry about why we pick different p-value cut-offs for different oncogenes.)
3) Fill out the tables on the next page, using the Excel MATCH function to count the overlap between gene sets, and the dhyper function in R to compute the one-sided Fisher's exact tests (use **54614** as the reference population).
4) Determine enrichment of the IGF signature within the Bild Ras samples using Gene Set Enrichment Analysis (GSEA).
    a. For "Expression dataset," load the Bild dataset (click the "Load data" button to do this).
    b. For the "Gene sets database" field, construct a "GMX" file with two columns, one for your "IGF up" genes (Step 1) and one for your "IGF down" genes. (Load the GMX file into GSEA using the "Load data" button. Info on GMX format at http://www.broad.mit.edu/gsea/wiki/index.php/Data_formats.)
    c. For "number of permutations, select 100.
    d. For "Phenotype labels," upload the "Bild_classes.cls" file provided for you, then select the "RAS_versus_REST" option for the "Phenotype labels" field.
    e. Select "false" for "Collapse dataset to gene symbols."
    f. Select "gene_set" for "Permutation type."
    g. For "Chip platform(s)," select "HG_U133_Plus_2.chip."
    h. Click the "Run" button. When finished, fill in the results in Table 3 on next page.

Table 1. Number of transcripts in each gene expression signature.

	up	down
IGF		
Myc		
Ras		
E2F3		
beta-catinin		
Src		

Table 2. Transcripts intersecting between gene sets.

	Number intersecting	P-value*
IGF up X Myc up		
IGF up X Ras up		
IGF up X E2F3 up		
IGF up X Bcat up		
IGF up X Src up		
IGF down X Myc down		
IGF down X Ras down		
IGF down X E2F3 down		
IGF down X Bcat down		
IGF down X Src down		

* by one-sided Fisher's exact test.

Table 3. GSEA results.

Gene set enriched in phenotype RAS.							
Gene Set	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
Gene set enriched in phenotype REST.							
Gene Set	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX