Exercise 10: Gene Set Enrichment Analysis (GSEA)

May 2, 2017

You continue to work on the same *E. coli* overexpression dataset that you received a week ago. Now, it is clear that the 12 genes that have been overexpressed are the following:

- strain01 – `dinI`
- strain02 – `dinB`
- strain03 – `lexA`
- strain04 – `rydC`
- strain05 – `luc`
- strain06 – `mazF`
- strain07 – `recA`
- strain08 – `relA`
- strain09 – `ruvA`
- strain10 – `sulA`
- strain11 – `umuD`
- strain12 – `uvrA`

You read a bit online about each of these genes, and find out that *mazF* is part of a very interesting toxin-antitoxin system. You are curious which known transcription factors are related to the function of this gene.

1 First look at the data

1.1 Volcano plot

Make a volcano plot comparing the expression in strain06 with strain01. Don't forget to correct for multiple hypotheses. Assume the same thresholds for fold-change and p-value as before (2 and 0.05). Do you notice anything unusual?

Hint: Follow the instructions in exercise 9.

1.2 Regulation data

Load the provided MATLAB file (`ex10.mat`), which contains two variables:

- `transcription_factors` – a cellarray with the names of 160 *E. coli* transcription factors.
- `regulation` – a matrix containing the genetic regulation coefficients. The rows correspond the genes in `genes` and the columns correspond to the 160 transcription factors (TFs). `regulation(i, j)` $= 1$ means that gene i is induced by TF j. -1 means the gene is repressed. 0 means no genetic interaction at all.

Which genes are repressed by *lacI*? Are there any genes that are induced by it?

Hint: First, find the index of the TF, using `find(strcmp(transcription_factors, 'lacI'))`. Then, use `find()` again to see which values in that column in `regulation` are positive, and which are negative. Finally, find the corresponding gene names in `genes`.

2 Enrichment Analysis

Since there are so many genes that were up/down regulated by *mazF*, you decide to perform a Gene Set Enrichment Analysis to see which transcription factors might be responsible for these changes.

2.1 Get the vector of features ranked by p-value for the TF *purR*

Generate a vector containing the features (i.e. whether a gene is regulated by *purR* or not) ordered by decreasing p-values from the previous question. Exclude those genes whose log2FC is smaller than 0.2.

Hint: Find the column that corresponds to *purR*:

```
col = find(ismember(transcription_factors, ’purR’));
```

Get a sorted list of gene indices (`I`):

```
[~, I] = sort(PValues + (abs(log2FC) < 0.2));
```

Note that the added term `(abs(log2FC) < 0.2)` is a trick to lower the rank of genes with low fold-change to be always at the bottom of the list. Finally, select the features from the regulation matrix and reorder them according to `I`:

```
features = (regulation(I, col) ~= 0);
```

2.2 Build contingency table for the gene set of size 50

How many genes are both significant (S+) and in group (G+)? Calculate all four values in the contingency table: A, B, C, and D.

Hint:

	In group (G+)	Not in group (G-)
Significant (S+)	A = sum(features(1:50));	B = sum(~features(1:50));
Not significant (S-)	C = sum(features(51:end));	D = sum(~features(51:end));

2.3 Hypergeometric test

For the contingency table you made for the gene set of size 50, calculate the probability of the null hypothesis according to Fisher’s Exact Test (or an hypergeometric distribution).

Hint: Choose one of the three equivalent options presented in class:

```
>> [~,p] = fishertest([A B;C D],’Tail’,’right’)
or >> p = sum(hygepdf(A:A+B, A+B+C+D, A+C, A+B))
or >> p = 1-hygecdf(A-1, A+B+C+D, A+C, A+B)
```

2.4 Build all contingency tables and calculate *p*-values

Using a loop, calculate the *p*-values for all gene set sizes between 2 and 1000.

2.5 Plot GSEA curve and find the lowest *p*-value (step 5)

Plot the result from the previous question as a function of the set size. Transform the *p*-values to $-\log_{10}$, to make it more readable. What is the size of the set with the smallest *p*-value and what is that *p*-value?

2.6 Repeat GSEA for all transcription factors

Calculate the minimal p-value for all 160 transcription factors (i.e. each one of the columns in the `regulation` matrix, store the results in a new matrix called `gsea_PValues`. Now, since each one of these p-values represents a test for a different hypothesis, you need to correct for multiple hypotheses. As always, use Storey's positive FDR method to do the correction.

Which transcription factor target sets are significantly enriched ($p < 0.05$) in strain06 compared to strain01?

Hint: There are 25 such TFs. Note that the loop might take a few minutes to run.