Gal Goldman

Gal.goldman@weizmann.ac.il

Finding common factors among pathways

**Introduction**

Biological pathway - series of actions that leads to a certain product or a change in a cell. Usually, multiple genes participate in a single biological pathway. In cancer, the operation of many important pathways is altered.

The Pathifier- calculates independently for every pathway, a score that represents the extent to which the pathway is deregulated in every individual sample. The calculations are based on the deviation of the sample from normal behavior as can be measured using the expression levels of genes belonging to each pathway (thus requiring previous knowledge to assign genes to pathways). The result is a PDS- pathway deregulation score. This allows high-level analyzing, that incorporates previous knowledge instead of ignoring it and working with thousands of genes in high dimensionality manner.

Databases- the Pathifier uses 624 pathways, taken from 3 databases: Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCarta and the National Cancer Institute–Nature Pathway Interaction Database (PID).

**Project’s goal**

Finding the common factors among the pathways in each cluster (defined by the Pathifier), in order to name the cluster in the optimal and most accurate way.

**Inputs:**

1. Pathifier output: an excel file in the following format:

* First row contains headlines.
* First column contains the pathifier names of the pathways:

1. For KEGG starts with the database name twice (KEGGKEGG) only big letters, words separated by '\_', and no additional punctuation signs.
2. For BIOCARTA starts with the database name (BIOCARTA) only big letters, words separated by '\_', and no additional punctuation signs.
3. For PID only small letters, words separated by '\_', and no additional punctuation signs.

* The second column is optional. If exists needs to be the division into clusters (‘A’/’B’ etc.).

1. The wanted cluster as a string. If you want to use the entire excel without using the second (and optional column), inset ‘all’.
2. Optional- your own pathways- this tool uses by default pathways' descriptions taken from 3 databases: KEGG, BIOCARTA, and PID. If you wish to insert your own pathways please write a path to an excel file in the following format:

* First row: headlines.
* First column: the name of the pathways without spaces or punctuation (in a way that would be a legal variable name according to matlab).
* Second column: an elaborated name for each pathway.
* Third column: optional- immune flag. If the pathway is immune system related the value would be 1, if not 0. Notice, if you wish to skip this option please insert 0 at all cells of this column.
* Fourth column and all the following columns: genes associated with the pathway.

**Outputs:**

The function prints to the command window the following statements:

1. The most common gene in the cluster's pathways, followed by a table that specifies additional genes that appeared in high frequency.
2. The most common word in the pathways' names, followed by a table that specifies additional words that appeared in high frequency.
3. A table of the most comprehensive superpathways- including the P-value for their overlap with the cluster.

The full tables (contains the ranks for all the genes, words and superpathways) are being saved to the user's computer, in the current folder (notice the current folder may change while choosing the input files), under the name: “cluster\_name\_output.xlsx” (genes table- sheet 1, words table- sheet 2, superpathways table- sheet 3).

**Main steps**

1. **Extract the needed files from inner directories:**

* Pathifier\_data\_file- only if the user didn’t insert a data file of his own.
* superpathways\_file
* Only genes of superpathways.

1. **Extract input from user:**Create a struct of genes per pathway for the relevant pathways.
2. **Immune score:**Count the number of pathways that has 1 in the third column- those are the immune related pathways. Create a message that would be printed later on with the information about the connection to the immune system or to diseases. In there aren’t any immune related pathways the message would not be printed.

1. **Common genes:**

Check whether there are genes that repeat in a substantial amount of the pathways.

Steps:

* Create a list with unique values only of all the genes that are on our pathways of interest.
* Create a matrix that represents the presence of genes in pathways. Matrix lines are the genes, and the columns are the pathways. If the gene in row i appears in column j the matrix value in place (i,j) would be 1, otherwise 0. Print to the command window the most common gene in the cluster (appeared at the highest number of pathways). If there are multiple genes that appear in the maximum value the function prints all of them. If there are any genes which the percentages of their appearance are smaller than the maximal gene by 15% or less, the function presents them too (the cutoff can be changed).
* Create a table of all the genes and their scores (number of pathways each gene appeared in, and the percentage of the pathways it appeared in out of all the relevant pathways). Print the first few lines.

1. **Characteristic words:**

For every pathway there are two names- the pathifier name (in some cases includes the database name, shorter, no spaces etc.), and an additional more official name. I wanted to identify repeating motifs in the pathways by checking if certain words repeat themselves more in the pathways’ names.

Steps:

* Separate name into individual words.
* Remove punctuation.
* Remove non indicative words (such as “and”, “of”, “by”, “"pathway", and numbers unattached to word etc.).
* Create a matrix similar to the one in the previous section, matrix lines are the words\names, and the columns are the pathways.
* Print to the command window the most common word in the cluster (appeared at the highest number of pathways). If the most common word is “signaling” print another word (since this word may be relevant but it appears a lot so it is highly non-indicative). In addition, If there are any words which the percentages of their appearance are smaller than the maximal word by 25% or less, the function presents them too (the cutoff can be changed).
* Create a table of all the words and their scores (number of pathways each word appeared in, and the percentage of the pathways it appeared in out of all the relevant pathways).

1. **Common superpathways:**

Steps:

* For every superpathway, count the number of genes (out of the pathways relevant to the cluster of course), that belong to that superpathway.
* Create a table of all the superpathways and their scores. The table contains the number of affiliated genes, number of total genes in this superpathway, and a P-value for the overlap between the genes in the cluster and the superpathway. The P-value was calculated using Or Zuk’s function that shows the probability to get a value as we got or higher. Since multiple comparisons are being conducted, the p-values that are presented are after FDR in the Benjamini and Hochberg method.

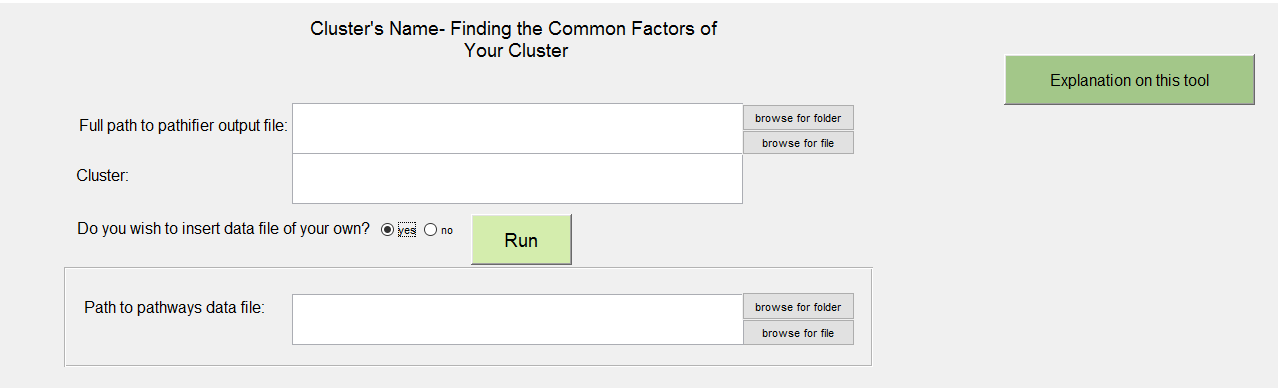
**Error messages:**

Error messages are presented in the following cases:

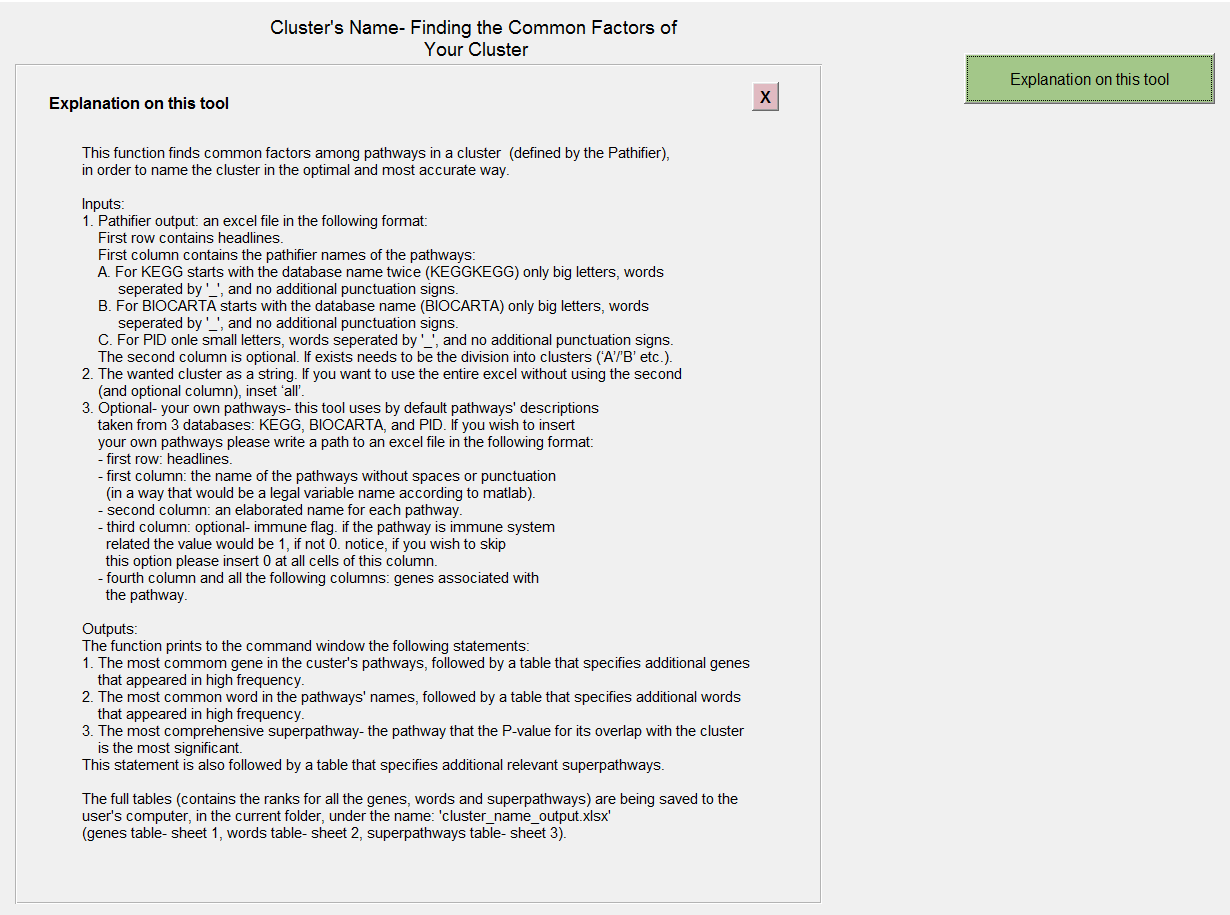
1. If the user tries to choose a file before a folder was chosen (GUI).
2. If the user writes a path to a file that doesn't exist (GUI).
3. If the specified cluster is not found in the input file.
4. If from any reason pathways were not chosen.
5. If the output file wasn't saved properly to the user's computer.
6. If the names of the pathways are not valid variables in matlab (the names in the pathifier format are valid).
7. If in the GUI graphical view the user asks for too many additional genes/words to be presented or if in request he inserts something which is not a number.

**Other comment:**

* I used: warning(‘off’, 'all') at the beginning of the code. If the pathifier name is too long to be a proper variable name it is being truncated in the process. I did not recognize a spot in which that truncation is problematic (but when it happens it creates a lot of warnings.

**GUI:**

You can either write the path or browse in the computer

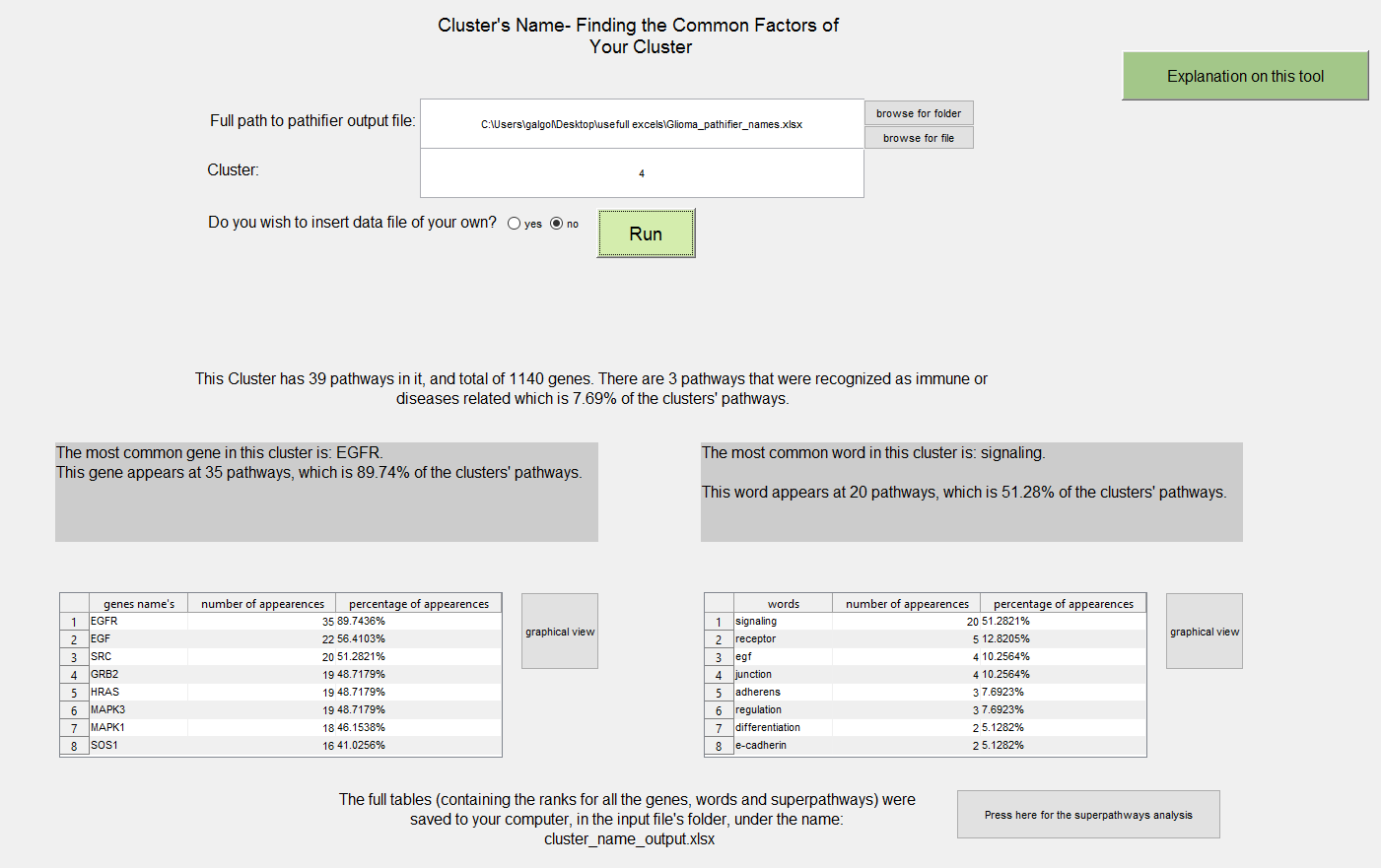
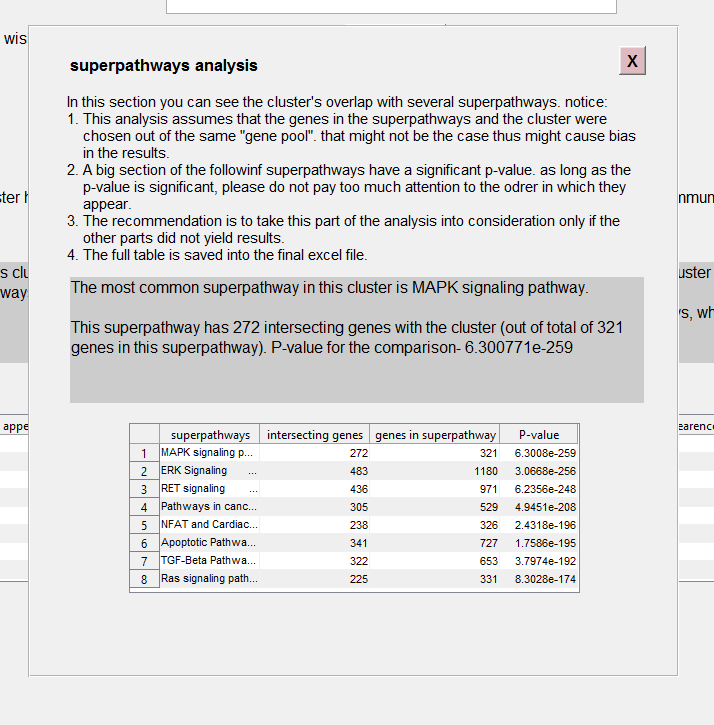


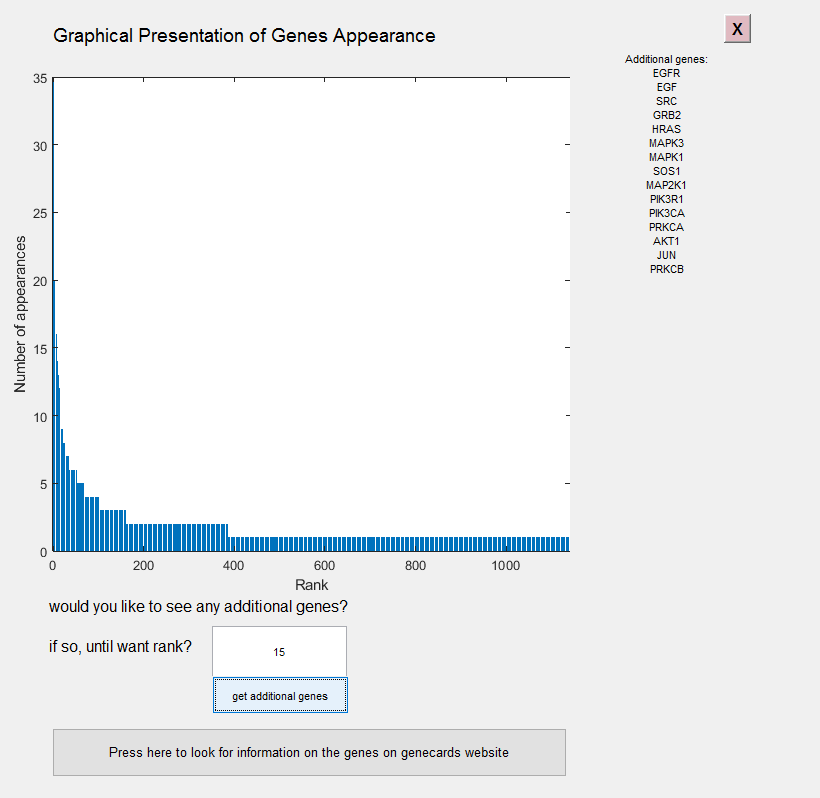
This window would open. Press on the X in order to close the window.

Press here for explanation on the tool

Graphical view: See the next page

General information + count of immune related pathways





Number of appearances vs rank

(An identical graph appears (if requested) for the common words)

Direct link to genecards website to search for more information on the genes.

Choose how many more genes\ words you would like to see and the list would appear aside from the graph.