Bioinformatics 2018-2019

Date: 04 December 2018

Project report

|  |
| --- |
| Bioinformatics@Data Science A.Y. 2018-2019  Manuscript Title  Nagham Almagout, Hassan Ismail and Alessandra Griesi  1Group no.9 Abstract In this project, we have a seed gene and we want to collect and store the basic information for all genes in this list, from HGNC database, and then for each seed gene, collect all binary protein interactions from Biogrid Human and IID Integrated Interactions Database, and finally, using innateDB, we find and report in tables the overrepresented GO categories and overrepresented pathways. |

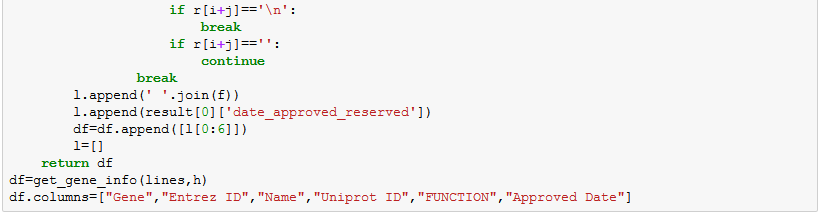
1. **Basic introduction about the disease/process**

Systemic lupus erythematosus (SLE) is a chronic disease that causes inflammation in connective tissues, such as cartilage and the lining of blood vessels, which provide strength and flexibility to structures throughout the body. The signs and symptoms of SLE vary among affected individuals, and can involve many organs and systems, including the skin, joints, kidneys, lungs, central nervous system, and blood-forming (hematopoietic) system. SLE is one of a large group of conditions called autoimmune disorders that occur when the immune system attacks the body's own tissues and organs.

SLE and other autoimmune disorders tend to run in families, but the inheritance pattern is usually unknown. People may inherit a gene variation that increases or decreases the risk of SLE, but in most cases do not inherit the condition itself. Not all people with SLE have a gene variation that increases the risk, and not all people with such a gene variation will develop the disorder, for example, a recent study showed synergistic interaction between risk alleles found in the following pairs of genes: HLA-CTLA4, IRF5-ITGAM, and PDCD1-IL21.

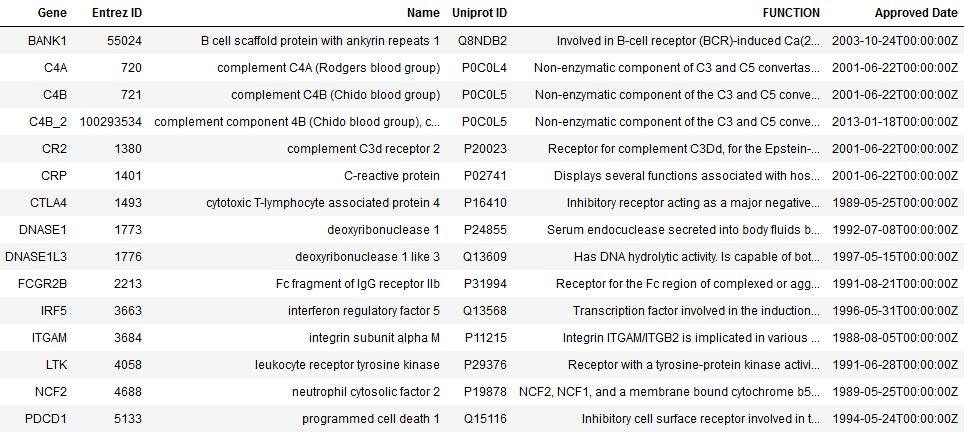
In rare cases, SLE can be inherited in an autosomal recessive pattern, which means both copies of the gene in each cell have mutations. The parents of an individual with an autosomal recessive condition each carry one copy of the mutated gene, but they typically do not show signs and symptoms of the condition.

1. **Seed genes**We used Python programming language to collect and store the basic information about seed gene, where we used bioservices, uniProt and pyhgnc libraries. By using these libraries, we was able to get: Entrez Gene ID, Protein Name, Uniprot AC, Protein Function and the last approved and updated date on HGCN.





And the result is:

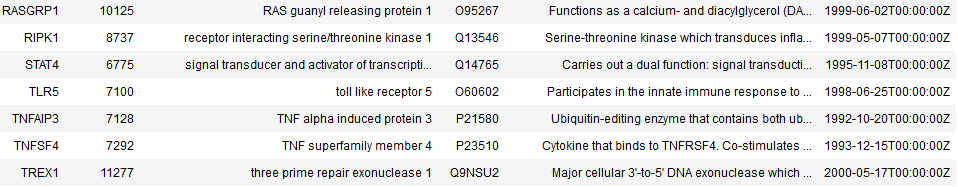


1. **Summary on interaction data**

Having the list of seed genes, we want to retrieve information about their interactions from two sources: (BioGrid, IID).

. BioGrid:

for BioGrid, the website provides a REST API service, which we can use its Python API as an interface to their database, in our code we contact the web service and request information about the interactions of a gene by providing gene symbol (from the seed gene list), the result is JSON format of the interaction of this gene, we do this procedure for all seed genes we have in our list, and we store these information in one dataframe and in a csv file.

After that we get the list of the interactors genes (from the previous interactions information), and we use them to request the interactions for each gene of them, and we save the new interactions of interactors in a dataframe and csv file.

IID:

for IID, first we were working on scrapping the information using python code, and we got some results, but unfortunately the page of the IID web service (http://iid.ophid.utoronto.ca) went offline few days ago , and they created a new page for the web service (http://178.128.224.72) but it was with some different HTML tags, so our code didn’t work anymore, that’s why we decided to go for the manual approach on retrieving the information we need.

(please note that we have checked the main page of IDD today and it was online again)

first we get the interaction information of the seed gene list, and the from it we get the interactors list.

### 3.2

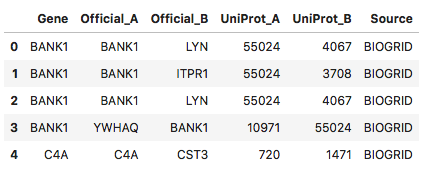
BioGrid:

Code of the function to retrieve the interaction information from BioGrid:



Sample of the table of interactions we retrieved from BioGrid:

in this table the UniProt information is not correct, that’s why we later mapped it to the correct UniProt values.

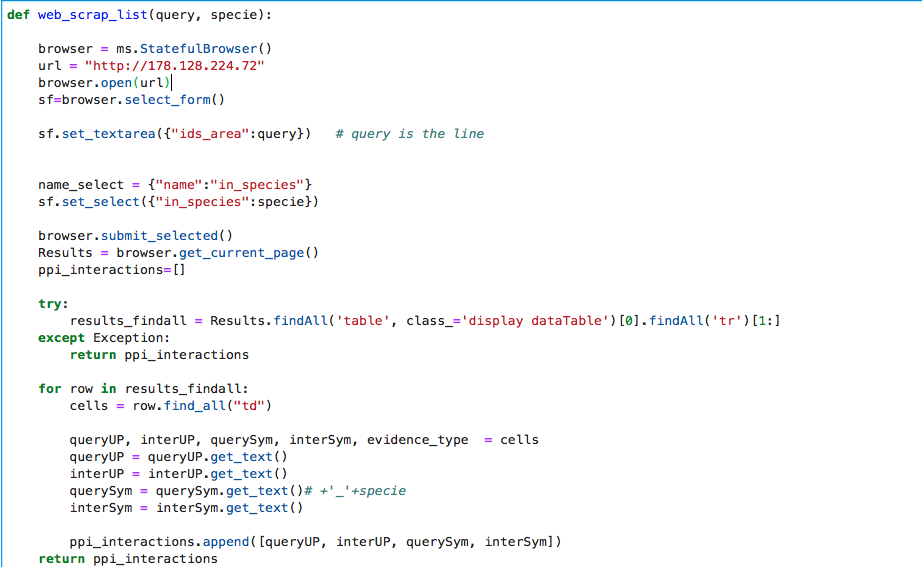


by the end of this step we collected the list of unique interactors with seed genes and stored them in a list: *BioGrid\_interactors*

Then we used the items of this list (interactors) to get information about the interaction among them using the below code:

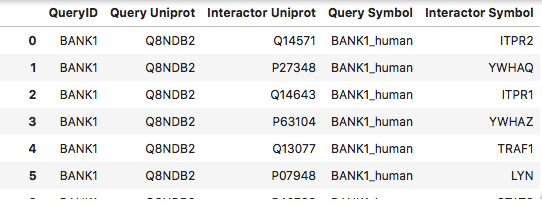


IID:  
the first code we used for the scrapping:s





and the result table was as below:



After that we went to do retrieving interactions of interactor manualy:

We saved the list of interactors of seed genes in a list (containing about 2200) and due to the limit of 100 item by query, we had to divide the list into 22 queries:

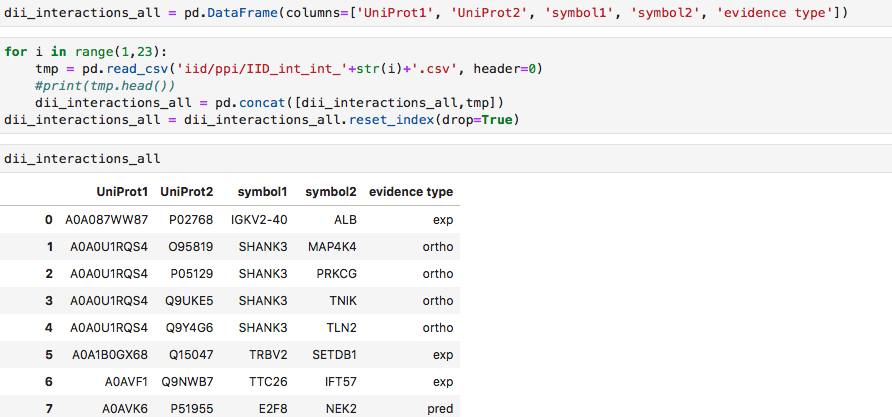
4.3

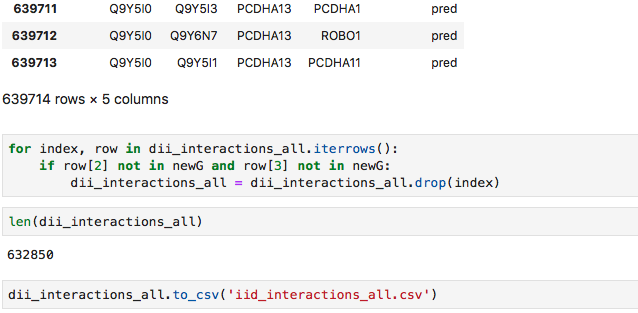
Please find the csv file (4.3\_intersections.csv) which contains the interactions found on both DBs.



and for each query we perform the search on the IID website and get a list of interactions.

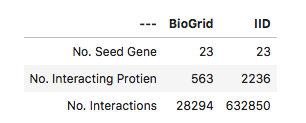
Then we combined all list of interactions into one matrix, keeping only interactions that include both interactors from the interactor list.





**3.3**

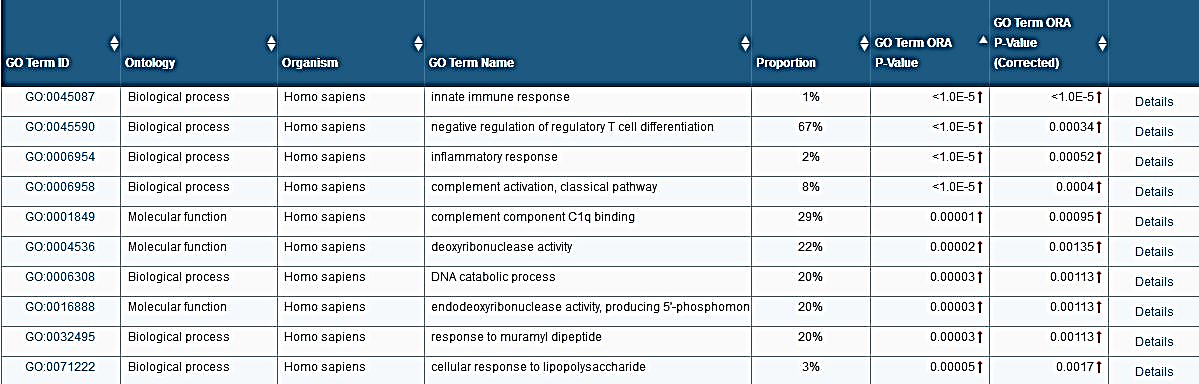
By comparing the results form the two sources:

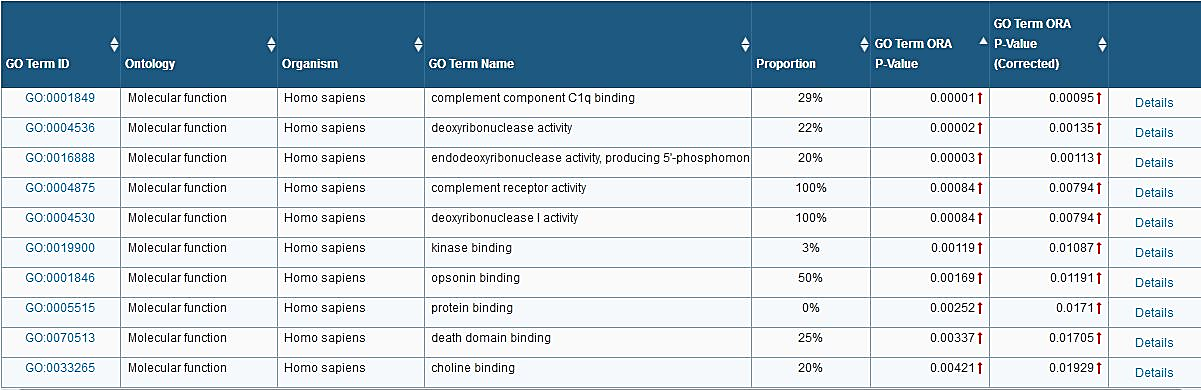


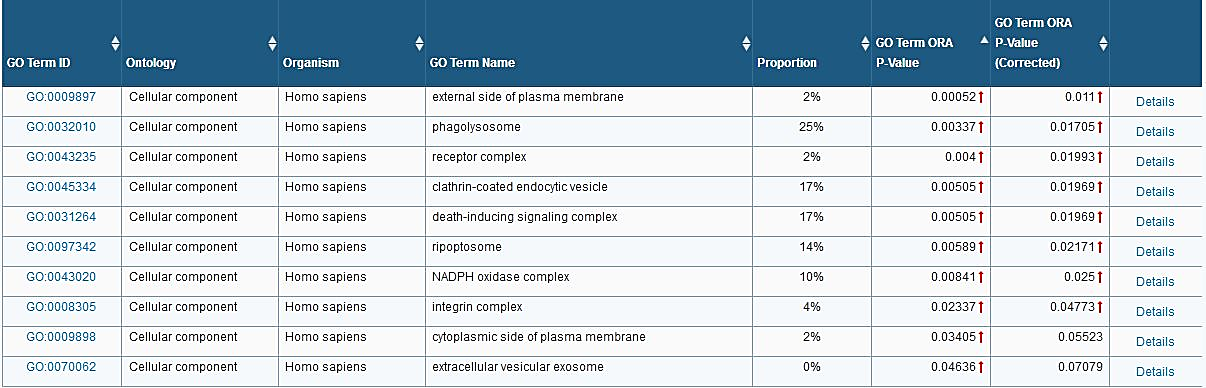
we can notice the big difference between the results, in term of Number of interactions of the same seed gene list, we got 563 interacting proteins (including the seed genes) from BioGrid, but we got 2236 interacting from IID, and for the number of interactions among interactors we have 28294 based on BioGrid, and 632850 based on IID.

# Enrichment analysis

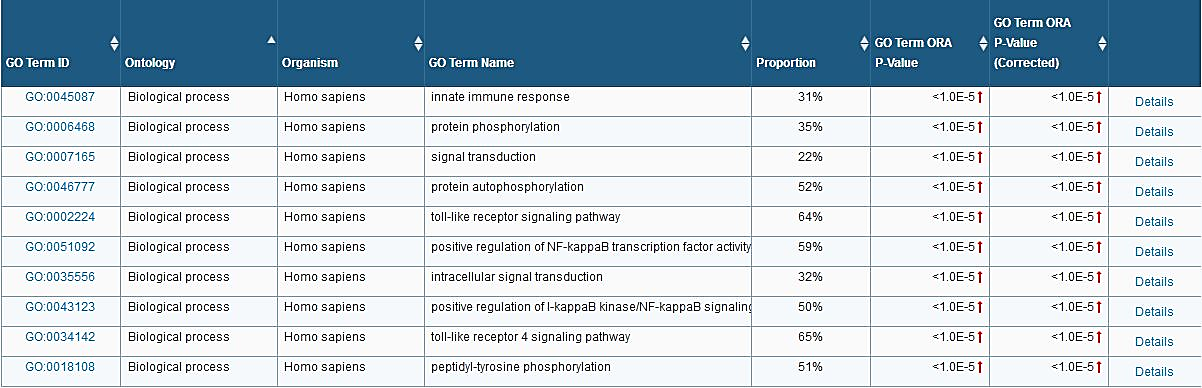
We used innateDB website manually to find the overrepresented GO categories and pathways for the seed gene interactome, union interactome, and intersection interactome. For this point, we used the Uniprot AC of mentioned interactome from the results that we got in points 3 and 4, to get the required information.

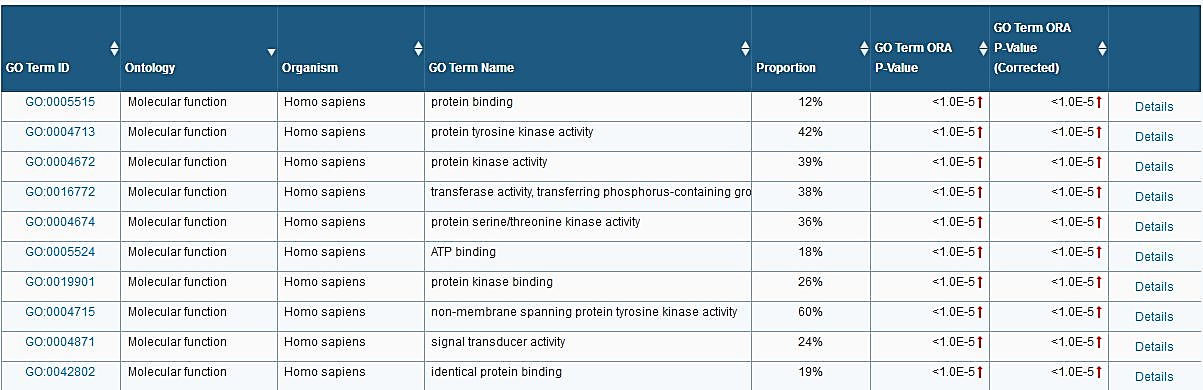
* + - 1. GO overrepresented:
* Gene seed

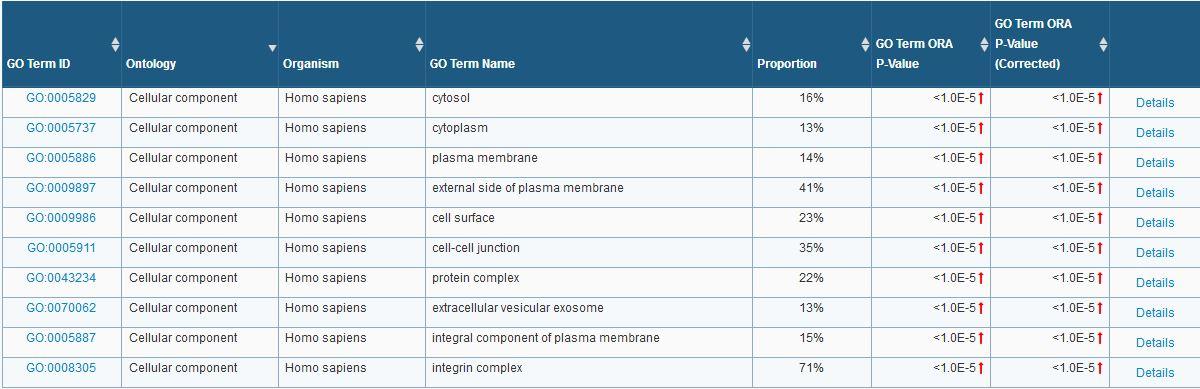




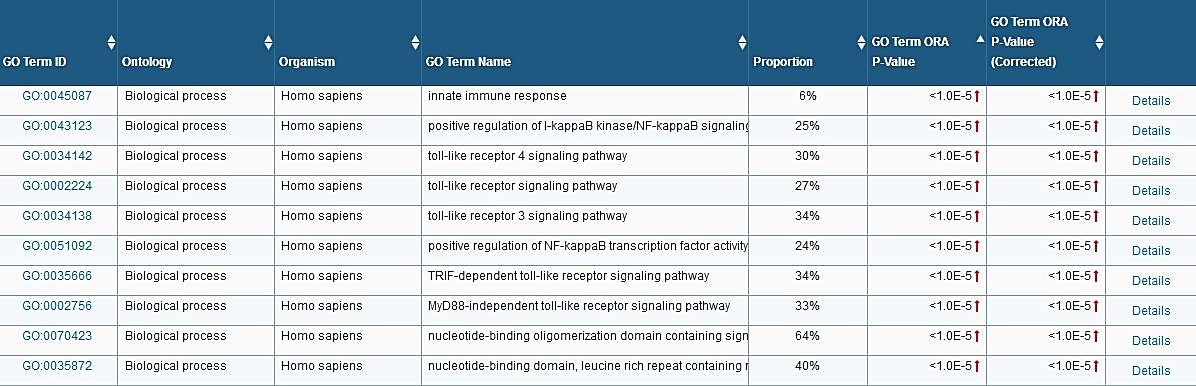
* Union interactome

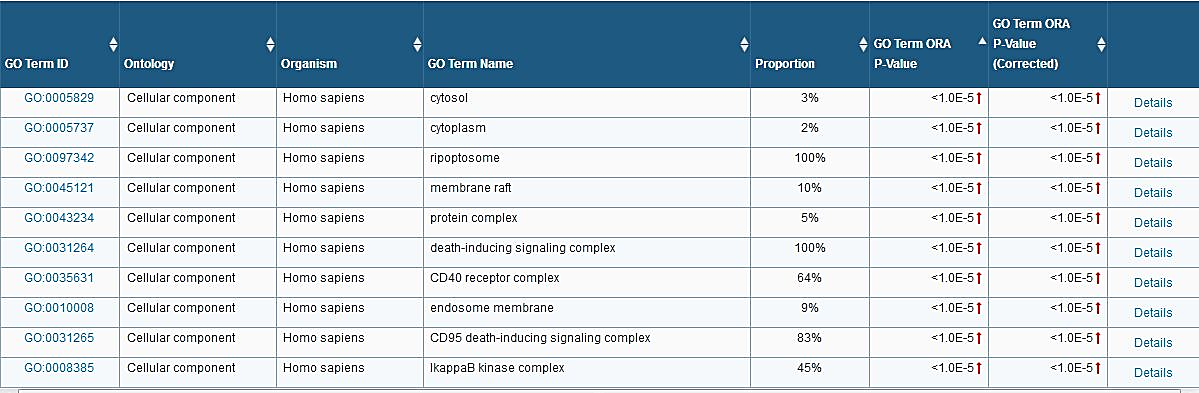
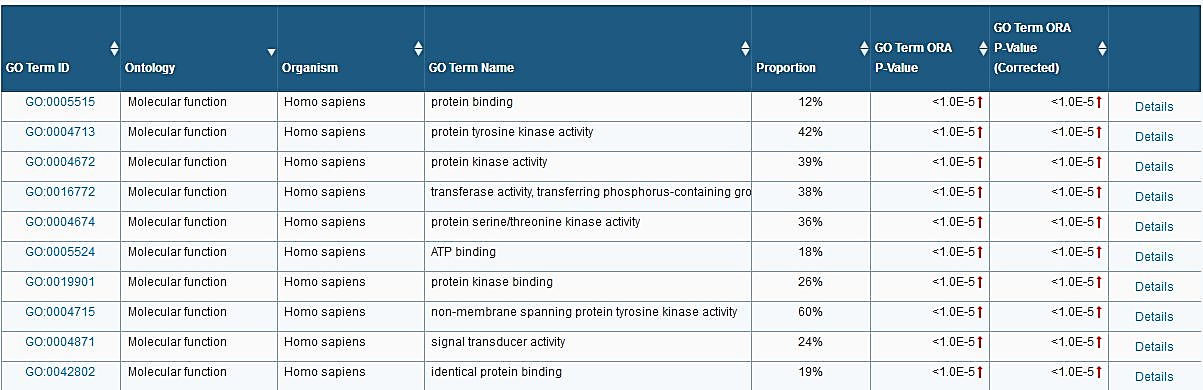


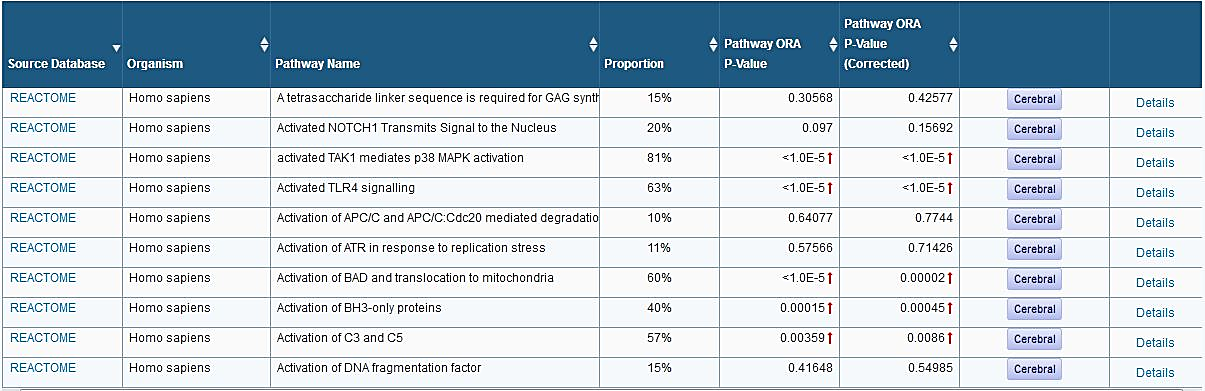
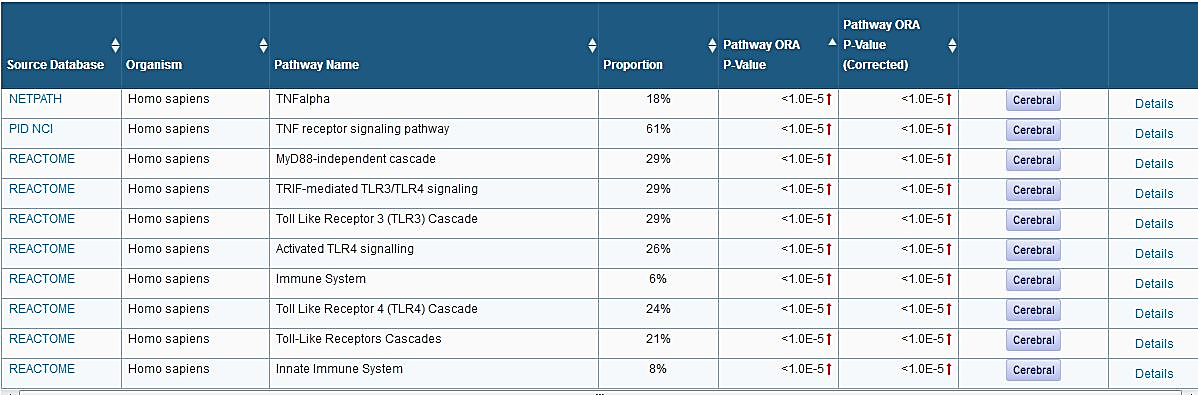
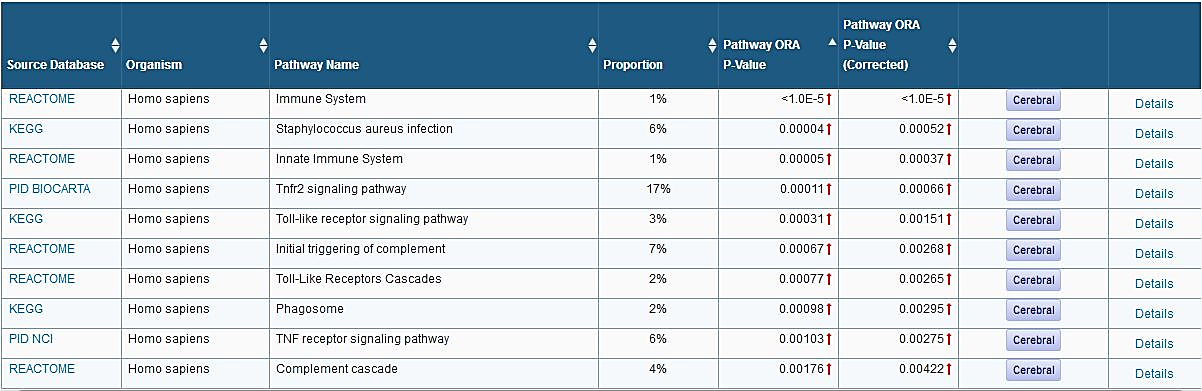




* Intersection interactome





* + - 1. Pathways overrepresented:
* Gene seed
* Union Interactome
* Intersection interactome