Bioinformatics 2018-2019

Date: 05 December 2018

Project report

|  |
| --- |
| Bioinformatics@Data Science A.Y. 2018-2019  Network Biology Project  Malick Alexandre Ngorovitch Sarr1, Pratuat Amatya1 and Viger Durand Azimedem Tsafack1  1Group no. 2 Abstract The Chagas disease is a parasitic disease caused by the protist Trypanosoma cruzi. Mainly found in South America, It is spread mostly by insects known as Triatominae, or "kissing bugs". In this project, we will be carring out data analysis of the human genes involved in the disease. The relevant set of genes was provided by the instructor after exploring (experiments and datasets, literature etc). We then proceeded in carrying out the various protein to protein interactions related to the seed genes for the Chaga disease. Finally, gene oncology was run in order to better understand the biological processes, molecular function and cellular of all the genes directly or indirectly related to the Chagas disease in humans. |

Basic introduction about the disease/process

Chagas disease, also known as American trypanosomiasis, is a tropical parasitic disease caused by the protist Trypanosoma cruzi.[1] The disease was named after the Brazilian physician and epidemiologist Carlos Chagas, who first described it in 1909. Chagas' disease is the most lethal endemic infectious ailment in the Western Hemisphere, with a devastating effect upon populations in rural areas of Latin America. Chagas' heart disease typically kills people in the age range of 30 to 50 years. The disease is considered incurable, and its high mortality rates translate to hundreds of thousands of deaths per year [2]. In Chagas-endemic areas, the main mode of transmission is through bite of an insect vector called a triatomine bug. The symptoms are transitive over the course of the infection. The symptoms may or may not be evident in early stages which includes fever, swollen lymph nodes, headaches or local swelling at the site of the bite. The chronic phase of disease starts after 8-12 weeks of infection during which 60-70% of the time no further symptoms are developed. The other 30–40% of people develop further symptoms 10–30 years after the initial infection, including enlargement of the ventricles of the heart in 20–30%, leading to heart failure.

Seed genes

With the available set of seed genes, we collected and stored the following basic information (summarized in the table 2.1.) using a python script (***/Source/2\_Collect\_data.py***) that makes use of the “**bioservices**” library to collect the information needed from the “**Uniprot”** platform and save them in local.

Table 2.1 Basic information about the seed gene

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Input gene name | HGNC Approved gene symbol | Uniprot AC | protein name | Status |
| BAT1 | SLC7A9 | P82251 | b-type amino acid transporter 1 | reviewed |
| CCL2 | CCL2 | P13500 | C-C motif chemokine 2 | reviewed |
| CCR5 | CCR5 | P51681 | C-C chemokine receptor type 5 | reviewed |
| CxCL10 | CXCL10 | P02778 | C-X-C motif chemokine 10 | reviewed |
| CXCL9 | CXCL9 | Q07325 | C-X-C motif chemokine 9 | reviewed |
| HLA-DPB1 | HLA-DPB1 | P04440 | HLA class II histocompatibility antigen, DP beta 1 chain | reviewed |
| HLA-DQB1 | HLA-DQB1 | P01920 | HLA class II histocompatibility antigen, DQ beta 1 chain | reviewed |
| HLA-DRB1 | HLA-DRB1 | Q29974 | HLA class II histocompatibility antigen, DRB1-16 beta chain | reviewed |
| IFNG | IFNG | P01579 | Interferon gamma | reviewed |
| IL10 | IL10RA | Q13651 | Interleukin-10 receptor subunit alpha | reviewed |
| IL12B | IL12B | P29460 | Interleukin-12 subunit beta | reviewed |
| IL1B | IL1B | P01584 | Interleukin-1 beta | reviewed |
| IL1RN | IL1RN | P18510 | Interleukin-1 receptor antagonist protein | reviewed |
| IL4 | IL4R | P24394 | Interleukin-4 receptor subunit alpha | reviewed |
| IL4R | IL4R | P24394 | Interleukin-4 receptor subunit alpha | reviewed |
| IL6 | IL6 | P05231 | Interleukin-6 | reviewed |
| LTA | LTA4H | P09960 | Leukotriene A-4 hydrolase | reviewed |
| TGFB | TGFB1 | P01137 | Transforming growth factor beta-1 proprotein | reviewed |
| TNF | TNF | P01375 | Tumor necrosis factor | reviewed |
| TNFA | TNF | P01375 | Tumor necrosis factor | reviewed |
| TNFB | LTA | P01374 | Lymphotoxin-alpha | reviewed |

The small summary of the data collected can be found on table 2.1 above. The full table is available in the files coming along this report at ***Data/Manually\_integrated\_basic\_info.csv***.

From the input gene name provided, we collected the official gene symbol which corresponds to the HGNC approved symbols which are unique symbols and names for human loci, including protein coding genes, RNA genes and pseudogenes, to allow unambiguous scientific communication. We are going to use those unique symbols as identifiers for computations in following sections.

We used The ***Uniprot AC*** as a unique identifier as well. It gave us access to a comprehensive resource for protein sequences and annotation data. For instance, starting From the *Uniprot AC*, it is possible to collect data such as the protein full name, its function within the human body and whether or not it has been reviewed at least once. Been reviewed for a protein means that it has been manually checked, annotated, reviewed and is present in the ***Switz-prots*** database. The *Switz-prots* database is a high quality annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions. Finally we collected the ***Entrez Gene ID*** (a.k.a GeneID) ***which*** is a unique integer used as a stable identifiers for genes and other loci for a subset of model organisms.

Summary on interaction data

After gathering the basic information about the input genes, we collected the Protein-to-Protein Interactions (PPI) using two PPI sources namely the *Biological General Repository for Interaction Datasets (****Biogrid)*** and the *Human Integrated Interactions Database* (IID).

## Biogrid Results

In order to extract the BioGRID interactions of interest, we downloaded the *BioGRID Tab 2.0 Delimited Text file format* of the data set in local (/*Data/BioGRID/BIOGRID-ORGANISM-Homo\_sapiens-3.5.167.tab2.txt)* and we ran it through a python script.

We used two identifiers to perform the search and retrieve the interactions of interest: ***GeneID*** and ***Gene Symbol***. The following are the steps we used to retrieve the interactions using both types of identifiers sequentially:

* Load both basic info table and full BioGRID PPI source data
* Filter our data by keeping all the rows having a seed gene as '**Entrez Gene Interactor A**' (or **Official Symbol Interactor A** if we are searching by gene official symbol). After this step, we will remain with the rows of interest
* Save all the protein-protein interactions from the above dataset
* Retrieve all the non-seed proteins interacting with at least one seed gene
* Retrieve and include in our interactions list the interactions among these non-seed proteins

More information and detailed comments about the implementation of these steps are available in the attached source code: ***Source/3\_Gather\_PPIs\_BioGRID.py***.

The python script produces the following output:

Performing the search of PPIs using BioGRID source...

Discrepancies observed between search by 'GeneIDs' an search by 'gene symbols':

Search by GeneID: 7309 interactions

Search by gene symbol: 7330 interactions

Summarize of the results:

Number of seed genes found: 16

Total Number of interacting proteins: 579

Total Number of interactions found: 7330

We found 16 out of 21 seed genes in the BioGRID using both GeneID and Gene symbols as search filters. Starting from those 16 genes, we found 579 interacting proteins including the non-seed genes. Finally, we found 7730 interactions among all of them. Additionally some discrepancies were found as we have recorded 7309 interactions using the GeneID to filter the data and 7330 interactions using the Gene official Symbols. This may suggest that more than one symbol may point to the same Gene ID.

## IDD Results

Similarly, as we did with BioGRID, we downloaded the *IID Tab Delimited Text file format* of the data set in local (*/Data/IID/human\_annotated\_PPIs.txt*) and we ran it through a python script.

In this case, we used two type of identifiers in order to retrieve the needed interactions: **Uniprot AC** and **Gene Symbols**. The following are the steps we used to retrieve the interactions using both types of identifiers sequentially:

* Load the basic info table
* Load by chunks of 20000 rows the full IDD PPI source data. Loading the data at once caused a memory error
* Filter our data by keeping all the rows having a seed gene as ‘**Uniprot Interactor A**' (or **Official Symbol Interactor A** if we are searching by gene official symbol) keeping ‘Uniprot Interactor B’ unfiltered; repeat the process filtering B keeping A unfiltered and merge the two data sets. After this step, we will remain with the rows of interest
* Save all the protein-protein interactions from the above dataset
* Retrieve all the non-seed proteins interacting with at least one seed gene
* Retrieve and include in our interactions list the interactions among these non-seed proteins

More information and detailed comments about the implementation of these steps are available in the attached source code: ***Source/3\_Gather\_PPIs\_IDD.py***.

The python script produces the following output:

Performing the search of PPIs using IID source...

Discrepancies observed between search by 'Uniprot AC IDs' an search by 'gene symbols':

Search by Uniprot AC ID: 13115 interactions

Search by gene symbol: 3386 interactions

Summarize of the results:

Number of seed genes found: 18

Total Number of interacting proteins: 5380

Total Number of interactions found: 13115

In this case, we found 18 out of 21 seed genes in the IID using both **Uniprot AC** and **Gene symbols** as search filters. Starting from those 18 genes, we found interacting proteins including the non-seed genes. Finally, we found 13115 interactions among all of them. Additionally, we found some discrepancies as we recorded 13115 interactions using the Uniprot AC and only 3386 interactions using the Gene Symbols.

This is a table summarizing the results of this section (can be found in /Question 3/Summarize\_of\_PPI\_results.xlsx):

Table 3.1 Summaries of the PPI results

|  |  |  |  |
| --- | --- | --- | --- |
|  | Number of seed genes found | Total Number of interacting proteins | Total Number of interactions found |
| Biogrid | 16 | 579 | 7330 |
| IID | 18 | 5380 | 13115 |

# Interactomes data (*Find the code in /Source/4\_Arrange\_interaction\_data.py*)

The steps used to create the interactomes are straightforward. We downloaded a mapping that would be able to convert GeneIDs to Uniprot Ac as the BioGRID PPI does not contain a Uniprot AC. We loaded our various datasets: the basic info about the input genes (*Data/Manually\_integrated\_basic\_infos.csv*), the BioGRID PPI generated at question 3 (*Question 3/PPI\_BioGRID.csv*) and the IDD PPI generated at question 3 (*Question 3/IDD\_PPI*).

We started by creating a full interactome dataframe made of the two PPIs by appending one PPI dataframe to another (appending the IDD PPI to the BioGRID PPI):

* The **seed gene interactome** was extracted from the full interactome dataframe by looking up the rows in which the *Uniprot AC* (A and B) matches the *Uniprot AC* of our seed genes from the basic info dataframe
* The **union interactome** was extract from the full interactome dataframe by looking up the rows in which at least one seed gene Uniprot AC appears.
* The **intersection interactome** was obtained by merging both BioGRID and IDD PPIs, keeping the rows appearing in both databases in which at least one gene is a seed gene.

Table 4.1 Sample interaction among Seed genes (file in */Question 4/seed\_genes\_interactome.csv*)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| interactor A gene  symbol | interactor B gene  symbol | interactor A Uniprot AC | interactor B Uniprot AC | database source |
| IFNG | IFNG | P01579 | P01579 | BioGRID |
| CCL2 | CCL2 | P13500 | P13500 | BioGRID |
| TGFB1 | TNF | P01137 | P01375 | IID |
| TGFB1 | TGFB1 | P01137 | P01137 | IID |
| TGFB1 | IL6 | P01137 | P05231 | IID |

Table 4.2 Sample union interactome (file in */Question 4/union\_interactome.csv*)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| interactor A gene  symbol | interactor B gene  symbol | interactor A Uniprot AC | interactor B Uniprot AC | database source |
| CCL2 | ORC4 | P13500 | B7Z5F1 | BioGRID |
| CCL2 | ORC2 | P13500 | Q13204 | BioGRID |
| CCL2 | MCM7 | P13500 | Q96D34 | BioGRID |
| TGFB1 | CXCR4 | P01137 | P61073 | IID |
| TGFB1 | IGFBP3 | P01137 | P17936 | IID |

Table 4.3 Sample interesection interactome (file in /Question 4/intersection\_interactome.csv)

|  |  |  |  |
| --- | --- | --- | --- |
| interactor A gene  symbol | interactor B gene  symbol | interactor A Uniprot AC | interactor B Uniprot AC |
| IFNG | IFNGR2 | P01579 | P38484 |
| IFNG | IFNG | P01579 | P01579 |
| IFNG | GOPC | P01579 | Q9HD26 |
| IL12B | IL23A | P29460 | Q9NPF7 |
| LTA | LGALS2 | P01374 | P05162 |

# Enrichment analysis

Using InnateDB online GO (Gene Ontology) analysis service, we performed GO Enrichment analysis for above obtained gene sets, namely seed gene list, union interactome and intersection interactome. We used default Hypergeometric analysis algorithm alongside Benjamini Hochberg algorithm for P-value correction. Following are the obtained results of overrepresented GO categories and pathways.

#### Table 5.1 Overrepresented GO categories for seed gene list (file in */Question 5/Seed\_genes\_ORA\_GO\_categories.xlsx*)

| **No** | **Biological Process (BP)** | **Molecular Function (MP)** | **Cellular Component(CC)** |
| --- | --- | --- | --- |
| 1 | immune response | cytokine activity | extracellular space |
| 2 | cytokine-mediated signaling pathway | peptide antigen binding | external side of plasma membrane |
| 3 | response to lipopolysaccharide | CXCR3 chemokine receptor binding | extracellular region |
| 4 | cellular response to lipopolysaccharide | chemokine activity | cell surface |
| 5 | inflammatory response | interleukin-1 receptor binding | MHC class II protein complex |
| 6 | positive regulation of interferon-gamma production | receptor binding | integral component of lumenal side of endoplasmic reticulum membrane |
| 7 | positive regulation of calcidiol 1-monooxygenase activity | cytokine receptor activity | clathrin-coated endocytic vesicle membrane |
| 8 | defense response to protozoan | tumor necrosis factor receptor binding | ER to Golgi transport vesicle membrane |
| 9 | negative regulation of growth of symbiont in host | growth factor activity | transport vesicle membrane |
| 10 | positive regulation of T cell proliferation | protein binding | trans-Golgi network membrane |

### Table 5.2 Overrepresented GO categories for Union Interactome (file in */Question 5/Union\_interactome\_ORA\_GO\_categories.xlsx*)

| **No** | **Biological Process (BP)** | **Molecular Function (MP)** | **Cellular Component(CC)** |
| --- | --- | --- | --- |
| 1 | innate immune response | protein binding | extracellular space |
| 2 | immune response | cytokine activity | extracellular region |
| 3 | inflammatory response | growth factor activity | cell surface |
| 4 | signal transduction | receptor activity | external side of plasma membrane |
| 5 | cytokine-mediated signaling pathway | receptor binding | plasma membrane |
| 6 | positive regulation of cell proliferation | chemokine activity | integral component of plasma membrane |
| 7 | extracellular matrix organization | identical protein binding | extracellular vesicular exosome |
| 8 | positive regulation of transcription from RNA polymerase II promoter | integrin binding | extracellular matrix |
| 9 | cell-cell signaling | protein homodimerization activity | cytosol |
| 10 | angiogenesis | protein kinase binding | cytoplasm |

### Table 5.3 Overrepresented GO categories for Intersection Interactome (file in */Question 5/Intersection\_interactome\_ORA\_GO\_categories.xlsx*)

| **No** | **Biological Process (BP)** | **Molecular Function (MP)** | **Cellular Component(CC)** |
| --- | --- | --- | --- |
| 1 | positive regulation of tumor necrosis factor production | interleukin-23 complex | interleukin-23 receptor binding |
| 2 | negative regulation of growth of symbiont in host | extracellular space | cytokine activity |
| 3 | positive regulation of osteoclast differentiation | interleukin-12 complex | interleukin-12 alpha subunit binding |
| 4 | positive regulation of interleukin-12 production | C-fiber | interferon-gamma receptor activity |
| 5 | interleukin-23 complex | synapse | interferon-gamma receptor binding |
| 6 | interleukin-23 receptor binding | trans-Golgi network transport vesicle | CCR2 chemokine receptor binding |
| 7 | innate immune response | dendrite | galactoside binding |
| 8 | positive regulation of NK T cell proliferation | axon terminus | interleukin-12 receptor binding |
| 9 | regulation of tyrosine phosphorylation of Stat1 protein | rough endoplasmic reticulum | receptor binding |
| 10 | positive regulation of interferon-gamma production | endocytic vesicle | cytokine receptor binding |

### Table 5.3 Overrepresented pathways for the seed genes, the union Intereactome and the intersection interactome (file in */Question 5/ORA\_GO\_Pathways.xlsx*)

| **No** | **Seed Genes** | **Union Interactome** | **Intersection Interactome** |
| --- | --- | --- | --- |
| 1 | Cytokine-cytokine receptor interaction | Cytokine-cytokine receptor interaction | Cytokine-cytokine receptor interaction |
| 2 | IL23-mediated signaling events | Immune System | IL23-mediated signaling events |
| 3 | Type I diabetes mellitus | Cytokine Signaling in Immune system | Chagas disease (American trypanosomiasis) |
| 4 | IL27-mediated signaling events | JAK STAT pathway and regulation | Jak-STAT signaling pathway |
| 5 | Toxoplasmosis | Pathways in cancer | Type I diabetes mellitus |
| 6 | Leishmaniasis | Osteoclast differentiation | IFN gamma signaling |
| 7 | Graft-versus-host disease | Jak-STAT signaling pathway | Leishmaniasis |
| 8 | Malaria | Innate Immune System | JAK STAT pathway and regulation |
| 9 | Chagas disease (American trypanosomiasis) | TNFalpha | No2-dependent il-12 pathway in nk cells |
| 10 | African trypanosomiasis | Signaling by Interleukins | Regulation of IFNG signaling |

# Notes and comments

References (if any)

1. "Chagas disease (American trypanosomiasis) Fact sheet N°340". World Health Organization. March 2013. Archived from the original on 27 February 2014. Retrieved 23 February 2014.
2. Pathogenesis of Chagas' Disease: Parasite Persistence and Autoimmunity, Antonio R. L. Teixeira,\* Mariana M. Hecht, Maria C. Guimaro, Alessandro O. Sousa, and Nadjar Nitz
3. HUGO Gene Nomenclature Committee at the European Bioinformatics Institute", NIH, <https://www.genenames.org/about/>
4. The UniProt Consortium UniProt: the universal protein knowledgebase [Nucleic Acids Res. 45: D158-D169 (2017)](http://dx.doi.org/doi:10.1093/nar/gkw1099)
5. Entrez Gene: gene-centered information at NCBI. Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. 2011.
6. [Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. Biogrid: A General Repository for Interaction Datasets. Nucleic Acids Res. Jan 1, 2006; 34:D535-9](http://www.ncbi.nlm.nih.gov/pubmed/16381927).
7. <http://iid.ophid.utoronto.ca/#context_filter> Utoronto.ca