Bioinformatics 2017-2018
Date: 03 December 2017
Project report

Bioinformatics@Data Science A.Y. 2017-2018

Chagas Disease

MATADEEN Craig[1], SELEK Aydan[1] and VUKOVIC Vlado[1]

[1]Group no. 1

Abstract

In this project, we dealt with Chagas disease in order to understand which genes and what kind of interactions between genes lead to this disease. With the provided list of seed genes we collected the respective updated data associated to the seed genes and then delved into the interactions of these genes among themselves and with those of other proteins that result in a person getting infected with the disease. These interactions were curated and stored in tables for later use. From the interactions various sub sets of interactions between the data sources were obtained and from these secondary interaction datasets further analysis in the form of GO categories and pathway analysis was done. Further details of these genes and interactions are stated in the following sections. It is our hope that this study will inform practitioners about management practices for gaining control over Chagas disease.

1 Basic introduction about the disease/process

Chagas disease, also known as American trypanosomiasis, is a tropical parasitic disease caused by the prosit Trypanosoma cruzi. It is spread mostly by insects known as Triatominae or kissing bugs. The symptoms change over the course of the infection. In the early stage, symptoms are typically either not present or mild, and may include fever, swollen lymph nodes, headaches, or local swelling at the site of the bite. After 8–12 weeks, individuals enter the chronic phase of disease and in 60–70% it never produces further symptoms. The other 30 to 40% of people develop further symptoms 10 to 30 years after the initial infection, including enlargement of the ventricles of the heart in 20 to 30%, leading to heart failure. An enlarged esophagus or an enlarged colon may also occur in 10% of people. Trypanosoma cruzi is commonly spread to humans and other mammals by the blood-sucking "kissing bugs" of the subfamily Triatominae. The disease may also be spread through blood transfusion, organ transplantation, eating food contaminated with the parasites, and by vertical transmission (from a mother to her fetus).
Prevention mostly involves eliminating kissing bugs and avoiding their bites. Other preventive efforts include screening blood used for transfusions. A vaccine has not been developed as of 2017. Early infections are treatable with the medication benznidazole or nifurtimox. It is estimated that 6.6 million people, mostly in Mexico, Central America and South America, have Chagas disease as of 2015.

2 Seed genes

In order to check the official gene names we downloaded the entire HGNC database. With this database, we also retrieved some of the basic information such as Official Gene Symbol, Uniprot AC, Protein Name and Entrez Gene ID. Connecting to the database via Python we were able to fetch and

manipulate the data. We could not get the Uniprot ID and Function from the HGNC database, so we had to turn to Bioservices for Python to get the Uniprot ID. We were able to obtain the Uniprot ID by mapping directly from the Uniprot AC that we retrieved from the HGNC database. Function was not part of the Uniprot Bioservices' api, so we had to download and parse the Uniprot webpage associated to each of the Uniprot ACs. This was done utilizing requests and Beautifulsoup in Python. Finally, we ended up with the information as shown below (see Table 2.1):

TABLE 2.1

	Search Symbol	Official Symbol	Uniprot AC	Uniprot ID	Protein Name	Entrez Gene ID	Function	Notes
1	CCL2	CCL2	P13500	CCL2_HUMAN	C-C motif chemokine ligand 2	6347	Chemotactic factor that attracts monocytes and basophils but not neutrophils or eosinophils. Augments monocyte anti-tumor activity…	
2	CCR5	CCR5	P51681	CCR5_HUMAN	C-C motif chemokine receptor 5 (gene/pseudo-gene)	1234	Receptor for a number of inflammatory CC-chemokines including MIP-1-alpha, MIP-1-beta and RANTES…	
3	CxCL10	CXCL10	P02778	CXL10_HUMAN	C-X-C motif chemo-kine ligand 10	3627	Chemotactic for monocytes and T-lympho-cytes. Binds to CXCR3.	
4	CXCL9	CXCL9	Q07325	CXCL9_HUMAN	C-X-C motif chemo-kine ligand 9	4283	Cytokine that affects the growth, movement, or activation state of cells that participate in im-mune and inflammatory response.	
5	BAT1	DDX39B	Q13838	DX39B_HUMAN	DExD-box helicase 39B	7919	Involved in nuclear export of spliced and unspliced mRNA.	Official gene name changed
6	HLA-DPB1	HLA-DPB1	P04440	DPB1_HUMAN	major histocompati-bility complex, class II, DP beta 1	3115	Binds peptides derived from antigens that ac-cess the endocytic route of antigen present-ing cells (APC)…	
7	HLA-DQB1	HLA-DQB1	P01920	DQB1_HUMAN	major histocompati-bility complex, class II, DQ beta 1	3119	Binds peptides derived from antigens that ac-cess the endocytic route of antigen present-ing cells (APC)…	
8	HLA-DRB1	HLA-DRB1	P01911	2B1F_HUMAN	major histocompati-bility complex, class II, DR beta 1	3123	Binds peptides derived from antigens that ac-cess the endocytic route of antigen present-ing cells (APC)…	
9	IFNG	IFNG	P01579	IFNG_HUMAN	interferon gamma	3458	Produced by lymphocytes activated by specific antigens or mitogens.	
10	IL10	IL10	P22301	IL10_HUMAN	interleukin 10	3586	Inhibits the synthesis of a number of cytokines, including IFN-gamma, IL-2, IL-3, TNF and GM-CSF produced by…	
11	IL12B	IL12B	P29460	IL12B_HUMAN	interleukin 12B	3593	Cytokine that can act as a growth factor for ac-tivated T and NK cells, enhance the lytic ac-tivity of NK/lymphokine-activated	
12	IL1B	IL1B	P01584	IL1B_HUMAN	interleukin 1 beta	3553	Potent proinflammatory cytokine. Initially dis-covered as the major endogenous pyrogen, induces prostaglandin synthesis, neutrophil influx and activation, T-cell activation and	
13	IL1RN	IL1RN	P18510	IL1RA_HUMAN	interleukin 1 receptor antagonist	3557	Inhibits the activity of interleukin-1 by binding to receptor IL1R1 and preventing its associa-tion with the coreceptor IL1RAP for signal-ing.	
14	IL4	IL4	P05112	IL4_HUMAN	interleukin 4	3565	Participates in at least several B-cell activation processes as well as of other cell types (PubMed:3016727)…	
15	IL4R	IL4R	P24394	IL4RA_HUMAN	interleukin 4 receptor	3566	Receptor for both interleukin 4 and interleukin 13. Couples to the JAK1/2/3-STAT6 path-way…	
16	IL6	IL6	P05231	IL6_HUMAN	interleukin 6	3569	Cytokine with a wide variety of biological func-tions. It is a potent inducer of the acute phase response…	
17	LTA	LTA	P01374	TNFB_HUMAN	lymphotoxin alpha	4049	Cytokine that in its homotrimeric form binds to TNFRSF1A/TNFR1, TNFRSF1B/TNFBR and TNFRSF14/HVEM. In its heterotrimeric form with LTB binds to	LTA is the current symbol of TNFB

18	TNFB	LTA	P01374	TNFB_HUMAN	lymphotoxin alpha	4049	Cytokine that in its homotrimeric form binds to TNFRSF1A/TNFR1, TNFRSF1B/TNFBR and TNFRSF14/HVEM. In its heterotrimeric form with LTB binds to	Official gene name changed as LTA
19	TGFB	TGFB1	P01137	TGFB1_HUMAN	transforming growth factor beta 1	7040	Multifunctional protein that controls prolifera-tion, differentiation and other functions in many cell types...	Official gene name changed
20	TNF	TNF	P01375	TNFA_HUMAN	tumor necrosis factor	7124	Cytokine that binds to TNFRSF1A/TNFR1 and TNFRSF1B/TNFBR. It is mainly secreted by macrophages and can induce cell death of certain…	TNF is the current symbol of TNFB
21	TNFA	TNF	P01375	TNFA_HUMAN	tumor necrosis factor	7124	Cytokine that binds to TNFRSF1A/TNFR1 and TNFRSF1B/TNFBR. It is mainly secreted by macrophages and can induce cell death of certain…	Official gene name changed as TNF

3 Summary on interaction data

For each database (APID, STRING and BioGRID), we used the same method:
We downloaded the entire datasets associated to homo sapiens. Some datasets needed further cleaning and filtering to provide a fully homo sapiens dataset associated to the species number 9606. The dataset was loaded into a MySQL database, with relevant data structures and links between respective tables being created to speed up data retrieval. The dataset was linked back to the HGNC dataset to get the respective gene symbols. SQL queries were written to get the primary interactions, i.e. the interactions between seed genes and proteins. Using this primary interactions dataset, secondary interactions were found, i.e. the interactions of the seed gene interactors. Using these two interaction level datasets, the intermediary interactions were obtained. The intermediary interactions were combined with the primary interactions to obtain all the interactions for the seed genes associated to the database.

As a summary of this work, we found following number of information from each databases. (see Table 3.1)

TABLE 3.1

	APID	STRING	BioGRID
Total number of seed genes	19	18	18
Total number of interacting proteins (including seed genes)	239	4,547	428
Total number of interaction	1,402	2,002,964	5,200

4 Intersection interactome

We arranged the interactome data in the following three tables:

(1) Table 4.1 shows only the interactions between our seed genes.
(2) Table 4.2 shows the direct interactions between seed genes and the proteins for all databases.
(3) Table 4.3 shows the direct interactions between seed genes and proteins that appear in all three databases i.e. the intersection of 4.2 between the three data sources.

TABLE 4.1

Protein A Gene Symbol	Protein A Uniprot AC	Interaction Type	Protein B Gene Symbol	Protein B Uniprot AC	Database Source
TGFB1	P01137		IL4R	P24394	string
TGFB1	P01137	inhibition	IFNG	P01579	string
TGFB1	P01137		IL1RN	P18510	string
TGFB1	P01137	expression	TNF	P01375	string
LTA	P01374		CCL2	P13500	string

Here, we only show first 5 rows. (see “4_seed_genes_interactome.tsv” for full version)

TABLE 4.2

Protein A Gene Symbol	Protein A Uniprot AC	Interaction Type	Protein B Gene Symbol	Protein B Uniprot AC	Database Source
TGFB1	P01137		LCE2B	O14633	apid
TGFB1	P01137		CILP	O75339	apid
TGFB1	P01137		COL2A1	P02458	apid
TGFB1	P01137		DCN	P07585	apid
TGFB1	P01137		THBS1	P07996	apid

Here, we only show first 5 rows. (see “4_union_interactome.tsv” for full version)

TABLE 4.3

Protein A Gene Symbol	Protein A Uniprot AC	Protein B Gene Symbol	Protein B Uniprot AC
TGFB1	P01137	COL2A1	P02458
TGFB1	P01137	DCN	P07585
TGFB1	P01137	THBS1	P07996
TGFB1	P01137	MMP2	P08253
TGFB1	P01137	CCL5	P13501

Here, we only show first 5 rows. (see “4_intersection_interactome.tsv” for full version)

5 Enrichment analysis

For enrichment analysis, we chose InnateDB to work with. With this database, we applied two kinds of analysis: over-represented GO analysis and over-represented pathway analysis.

Manually by following the web-site procedure, we uploaded the list of genes (seed gene, union and intersection interactome gene list). In order to find the corresponding ID of genes in InnateDB, we made a cross-reference ID with Uniprot ID. We then applied ORA with recommended algorithm Hypergeometric for analysis and with recommended P-Value correlation method Benjamini Hochberg.

Once we got the all results, we sorted them by their p-value corrected value and we only took the first 10 results to create the tables.

Over-represented GO Analysis:

a) Seed genes GO analysis (see Table 5.1)
b) Union interactome GO analysis (see Table 5.2)
c) Intersection interactome GO analysis (see Table 5.3)

Over-represented Pathway Analysis:

a) Seed genes pathway analysis (see Table 5.4)
b) Union interactome pathway analysis (see Table 5.5)
c) Intersection interactome pathway analysis (see Table 5.6)

TABLE 5.1: Seed genes GO analysis

Pathway Name	Pathway ID	Source Name	Gene Count	Genes for this entity	Pathway p-value	Pathway p-value (corrected)
immune response	GO:0006955	biological process	15	367	1.1640796500077883e-23	8.544344631057167e-21
cytokine activity	GO:0005125	molecular function	10	171	1.1197265585600232e-16	4.1093964699152846e-14
cellular response to lipopolysaccharide	GO:0071222	biological process	8	87	5.944753731294622e-15	1.4544830795900844e-12
extracellular space	GO:0005615	cellular component	14	1144	1.9068104641381045e-14	3.4989972016934212e-12
response to lipopolysaccharide	GO:0032496	biological process	8	154	6.42890088451672e-13	9.437626498470544e-11
defense response to protozoan	GO:0042832	biological process	5	16	1.8200248957571937e-12	2.2264971224763003e-10
negative regulation of growth of symbiont in host	GO:0044130	biological process	5	17	2.576894910039251e-12	2.7020583770983003e-10
external side of plasma membrane	GO:0009897	cellular component	8	192	3.825501292998238e-12	2.807917949060707e-10
inflammatory response	GO:0006954	biological process	9	315	3.660944971217983e-12	2.9857040098599987e-10
response to drug	GO:0042493	biological process	9	313	3.4574068405887018e-12	3.1721707762401337e-10

TABLE 5.2: Union interactome GO analysis

Pathway Name	Pathway ID	Source Name	Gene Count	Genes for this entity	Pathway p-value	Pathway p-value (corrected)
innate immune response	GO:0045087	biological process	831	1384	4.680197287881685e-195	4.9488406122060936e-191
protein binding	GO:0005515	molecular function	3090	9632	8.205715333583185e-154	4.338361696865431e-150
cytosol	GO:0005829	cellular component	1138	2642	1.629805022896171e-120	5.744519437368038e-117
cytokine-mediated signaling pathway	GO:0019221	biological process	222	249	7.439847056759957e-107	1.9667235694544939e-103
poly(A) RNA binding	GO:0044822	molecular function	590	1104	2.0322599920442596e-104	4.2978234311752016e-101
ATP binding	GO:0005524	molecular function	721	1493	2.0721989017919196e-100	3.6519051979246246e-97
nucleoplasm	GO:0005654	cellular component	573	1132	9.493937722291228e-89	1.4341271067929637e-85
gene expression	GO:0010467	biological process	402	680	6.430646606495277e-88	8.499707152135127e-85
inflammatory response	GO:0006954	biological process	236	315	6.227033895488876e-82	7.316072934544377e-79
viral process	GO:0016032	biological process	336	545	5.3393348482067815e-80	5.645812668493851e-77

TABLE 5.3: Intersection interactome GO analysis

Pathway Name	Pathway ID	Source Name	Gene Count	Genes for this entity	Pathway p-value	Pathway p-value (corrected)
innate immune response	GO:0045087	biological process	63	1384	4.7668936142527475e-37	9.352645271163887e-34
immune response	GO:0006955	biological process	33	367	9.761765305027037e-28	9.576291764231524e-25
cytokine-mediated signaling pathway	GO:0019221	biological process	27	249	6.823854971413083e-25	4.462801151304156e-22
extracellular space	GO:0005615	cellular component	42	1144	4.203780622715898e-20	2.061954395442148e-17
inflammatory response	GO:0006954	biological process	25	315	9.994535235873571e-20	3.92185562655679e-17
protein binding	GO:0005515	molecular function	114	9632	3.765608581757017e-19	1.2313540062345448e-16
intronless viral mRNA export from host nucleus	GO:0046784	biological process	8	8	3.918976057620765e-18	1.0984330035788488e-15
apoptotic process	GO:0006915	biological process	30	639	3.3851909941106836e-17	8.302180913056453e-15
external side of plasma membrane	GO:0009897	cellular component	18	192	1.0585742921472634e-15	2.3076919568810333e-13
cell surface	GO:0009986	cellular component	24	432	1.9958181137552257e-15	3.915795139187752e-13

TABLE 5.4: Seed genes pathway analysis

Pathway Name	Pathway ID	Source Name	Gene Count	Genes for this entity	Pathway p-value	Pathway p-value (corrected)
Cytokine-cytokine receptor interaction	515	KEGG	14	258	4.378956489444901e-20	4.4665356192338e-18
Leishmaniasis	10355	KEGG	9	71	8.837276900031185e-16	4.507011219015903e-14
Allograft rejection	2793	KEGG	7	35	8.384096659759903e-14	2.8505928643183667e-12
IL23-mediated signaling events	15427	PID NCI	7	37	1.2812956118791391e-13	3.267303810291805e-12
Type I diabetes mellitus	525	KEGG	7	41	2.787742586804685e-13	5.686994877081558e-12
Malaria	10359	KEGG	7	49	1.0574141789781632e-12	1.7976041042628773e-11
IL27-mediated signaling events	15133	PID NCI	6	26	2.5262375983404012e-12	3.68108907186744e-11
Chagas disease (American trypanosomiasis)	10366	KEGG	8	103	2.9190984644921917e-12	3.7218505422275436e-11
Toxoplasmosis	10385	KEGG	8	119	9.50010386506997e-12	1.076678438041263e-10
African trypanosomiasis	10384	KEGG	6	35	1.7653709476052246e-11	1.8006783665573287e-10

TABLE 5.5: Union interactome pathway analysis

Pathway Name	Pathway ID	Source Name	Gene Count	Genes for this entity	Pathway p-value	Pathway p-value (corrected)
Immune System	18444	REACTOME	736	1127	9.070071059445121e-108	1.864806609821918e-104
Cytokine Signaling in Immune system	17418	REACTOME	247	267	1.92008196459492e-89	1.9738442596035782e-86
Cytokine-cytokine receptor interaction	515	KEGG	230	258	3.173816134026137e-75	2.1751219905192454e-72
G alpha (i) signalling events	13220	REACTOME	211	231	1.1168158311595881e-73	5.740433372160285e-71
GPCR ligand binding	19266	REACTOME	323	433	4.2123768133571834e-66	1.7321293456524737e-63
JAK STAT pathway and regulation	16125	INOH	226	273	8.180095479199715e-61	2.8030460508724355e-58
Interferon Signaling	18059	REACTOME	148	158	1.6822632201139084e-55	4.9410474007917065e-53
Adaptive Immune System	18371	REACTOME	388	604	2.1477952503399958e-51	5.519833793373788e-49
Chemokine signaling pathway	4389	KEGG	156	181	9.764440470097835e-47	2.2306321785023502e-44
Pathways in cancer	4397	KEGG	238	329	1.3358755885791087e-44	2.746560210118648e-42

TABLE 5.6: Intersection interactome pathway analysis

Pathway Name	Pathway ID	Source Name	Gene Count	Genes for this entity	Pathway p-value	Pathway p-value (corrected)
Cytokine-cytokine receptor interaction	515	KEGG	33	258	6.789822359454857e-26	4.596709737350939e-23
Toxoplasmosis	10385	KEGG	24	119	1.0140571781309071e-23	3.432583547973121e-21
JAK STAT pathway and regulation	16125	INOH	31	273	1.060303825243883e-22	2.3927522989670298e-20
Leishmaniasis	10355	KEGG	19	71	1.9053957360422206e-21	3.2248822832514585e-19
Signaling by Interleukins	18744	REACTOME	20	110	6.612170642024438e-19	8.952879049301088e-17
Cytokine Signaling in Immune system	17418	REACTOME	27	267	1.9430441192165847e-18	2.1924014478493797e-16
IL4	15920	NETPATH	17	74	4.786413752373936e-18	4.629145871938792e-16
IL12-mediated signaling events	15185	PID NCI	15	57	5.258685662493965e-17	4.450162741885518e-15
Jak-STAT signaling pathway	568	KEGG	21	158	6.413543676071755e-17	4.824410076333976e-15
Chagas disease (American trypanosomiasis)	10366	KEGG	18	103	8.628685828857327e-17	5.84162030613641e-15

6 Notes and comments

- From the provided list of 21 seed genes, these resulted in a unique listing of 29 genes/Uniprot ids, as two of the genes represented the old names associated to their respective current gene symbol.
- Interaction type was only available for the STRING database. For this reason, the interactions obtained from the other two database sources has an empty interaction type column.
- See the .sql file for the queries that were run against the databases. These queries represent how the primary, secondary and intermediary datasets for each data source was generated and also shows the final aggregation query for collecting the interactions for the respective data source.
- Most of the interaction data retrieval was done utilizing SQL, tying back into HGNC to get the current gene symbol. Of the three data sources, the STRING database was very complex and difficult to manage. From a database perspective, a lot more can be done to make these datasets more normalized. Also, not that for the STRING database, A-B interactions with multiple interaction types, the interaction types were aggregated into a single column to prevent horizontal growth of the interaction list.

References

1. World Health Organization (2013) "Chagas disease (American trypanosomiasis) Fact sheet N°340".
2. Rassi A. Jr., Rassi A., Marcondes de Rezende J. (2012) "American trypanosomiasis (Chagas disease)". Infectious disease clinics of North America. 26 (2): 275–91.