Informatics Coursework Cover Sheet
2020/21

THE UNIVERSITY of EDINBURGH

Please fill in the information below, in relation to the coursework you are submitting.

Exam Number:	B177407
Course Name:	Bioinformatics 1
Coursework Number:	CW2
Tutor's Name:	Ian Simpson

Date:		Signature:	

1. Introduction

SFARI Gene is an evolving database for the autism research community. The database assigns every gene with a score reflecting the strength of the evidence linking it to the development of autism.[1]

This report is about a research on Autism literature, genes and gene network using the genes from SFARI database with the gene score 1 to 3.

2. Data and Methods

The main data used in this report is from Data set: *SFARI-Gene_genes_10-29-2020release_11-22-2020export.csv* which was provided from the course website, NCBI Entrez Database[2] and NCBI gene2go file[3].

Most of the works are done by using Python(Bio, pandas, ul, seaborn). Some tools also used in this report are: the PantherDB tool[4] and the STRING website[5].

3. Tasks and Results

3.1. Part 1 - Autism Literature

3.1.1. Task 1

In this task, I used the pandas value_counts() function to count for each gene-score. The bar chart below shows the number of genes in each gene-score group:

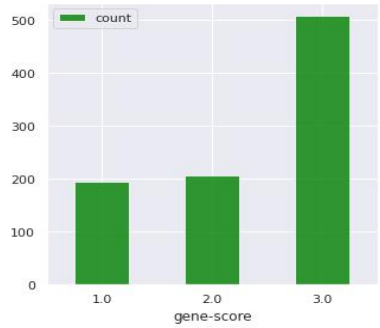

Figure 1: number of genes by different gene-score

3.1.2. Task 2

In this task, I just sorted the number of report and the top 5 SFARI genes with the most 'number-of-reports' ranked by different gene score is shown in table 1:

Gene Symbol	Number of Reports
SHANK3	88
NRXN1	88
MECP2	87
SCN2A	71
SCN1A	67

Table 1: top 5 SFARI genes with the most number of reports

3.1.3. Task 3

In this part, I retrieved data from PubMed using Entrez from Bio package.

The Medical Subject Headings (MeSH) thesaurus is a controlled and hierarchically-organized vocabulary produced by the National Library of Medicine.[6] Using the MeSH Major Topic is not just picking up papers that have autism as a word necessarily in it. It's picking up any paper that has been annotated with the mesh keywords that live within that part of the mesh annotation table.

So in this case, the keywords were set to *autism[MAJR] AND ("gene symbol" OR "gene name")* in which the [MAJR] represents MeSH Major Topic. Notice that I included not only the gene symbol but also the gene name because some articles refer to the gene as their name while others as their symbols.

The results are shown as follow:

Gene Symbol	Number of Articles
SHANK3	65
NRXN1	37
MECP2	80
SCN2A	12
SCN1A	11

Table 2: Number of Articles of each top 5 genes

3.1.4. Task 4

In this task, I used the keywords: *autism[MAJR] AND ("gene symbol" OR "gene name") AND "year"[DCOM]* in which the year ranges from 2000 to 2010 since these are the earliest and latest year of the appearance of articles about autism relating to the top 5 genes. As was discussed in the discussion forum, some paper may have different version or last for more than one year. To solve this problem, I just changed [DP] to [DCOM] which means the Completion Date pf the paper so that one paper is counted for only once. The sum of papers on each gene matches the total number of papers which can approve it.

Gene Symbol	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
SHANK3	0	0	0	0	0	0	0	2	5	9	0
NRXN1	0	0	0	0	1	0	0	0	4	5	1
MECP2	0	2	1	4	2	7	2	3	7	7	8
SCN2A	0	0	0	1	0	1	0	0	2	0	1
SCN1A	0	0	0	1	0	2	0	0	1	1	0

Gene Symbol	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
SHANK3	5	8	4	1	1	4	6	6	9	5
NRXN1	1	4	3	4	6	1	2	1	2	2
MECP2	2	5	5	5	2	7	5	5	0	1
SCN2A	0	1	1	1	0	1	0	0	3	0
SCN1A	0	2	1	0	0	0	1	0	1	1

Table 3: Number of Articles counted by year

3.1.5. Task 5

The single stacked histogram displaying the data from the table is shown below(Figure 2) as well as another single stacked bar chart(Figure 3) which will demonstrate the data clearer. The methods used were DataFrame.plot.hist() and DataFrame.plot.bar(). The analysis of this part is presented in section 4.1.

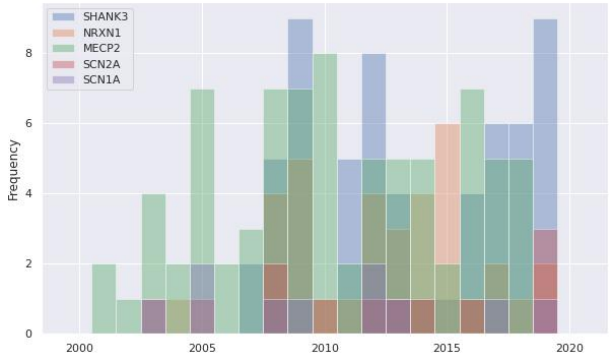

Figure 2: Number of Articles counted by year - histogram

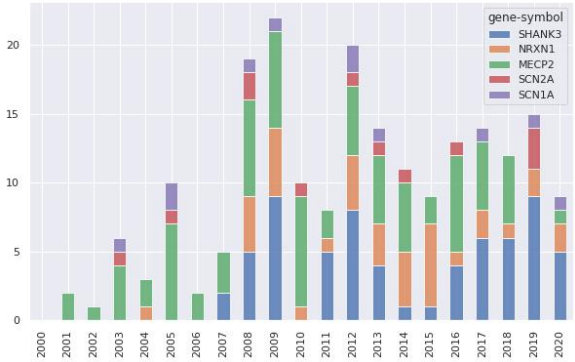

Figure 3: Number of Articles counted by year - bar chart

3.1.6. Additional Analysis

In this section, I did the PubMed search with the same key words as *autism[MAJR] AND ("gene symbol" OR "gene name") AND 0:2000[DCOM]* for all the SFARI genes in order to find those articles related to the genes which were published before **2000** since **2000** is the earliest year of the reports related to top **5** genes. Here are some significant results:

Gene Symbol	Number of Articles	Gene Symbol	Number of Articles
FMR1	7	*TEK*	3
GABRB3	7	*SON*	38
HRAS	3	*ADSL*	8
UBE3A	3	*GFAP*	3

Table 4: Number of Articles before 2000

I also did the PubMed search for every genes for the articles published recently: on **2019** and **2020.** And here are some significant results:

Gene Symbol	Counts 2019	Counts 2020
SHANK3	5	3
FMR1	5	2
CACNA1C	5	0
NRXN1	4	0
PTEN	2	3
NLGN3	2	1

Table 5: Number of Articles in 2019 and 2020

3.2. Part 2 - Autism Genes

3.2.1. Task 1

In this task, I used the ul package in python to retrieve the gene information of Homo Sapiens in order to link the gene symbols to the gene IDs. However, while doing so, there are some genes that are in the SFARI-Gene file but not in the gene-info file. Since there are only a few of them, after printing them out and searched manually, there are mainly **3** reasons:

1). The gene symbols in the SFARI-Gene file are sometimes aliases(pseudonym), for example: *FAM92B*(alias) in FSARI-genes is recorded as *CIBAR2* in NCBI database;

2). The gene symbols in the SFARI-Gene file actually refer to a group of genes, for example: gene *NOTCH2NL* is an alias for *NOTCH2NLA, NOTCH2NLB* and *NOTCH2NLC*;

3). The gene symbols in the SFARI-Gene file are too specific, for example: gene *RP11-1407O15.2* is recorded as *RP11* in NCBI database.

There is also a duplicated gene symbol in the data set: *MEMO1*. And after searching it on the NCBI website, there are indeed 2 IDs for this gene symbol. So I decided to keep both of them.

Finally, the total number of gene IDs mapped to the SFARI genes are 993.

3.2.2. Task 2

In this task, the same method was used to retrieve the file *gene2go.gz*. After merging gene2go file and the IDs of all the SFARI genes, a total 26969 hits are obtained.

However, there were 3304 rows that were fond duplicated: same GeneIDs with same GO_IDs.

After checking the gene2go file, I fond that this is because that there are many genes that hit the same GO term with different evidences. As a result, I decided to remove the duplicated rows since there should be only one hit of the same GO terms from one gene. After cleaning up the data, there are now 23665 hits.

3.2.3. Task 3

In this task, I split the genes with different gene-scores using groupby() and get_goup() methods in pandas.

3.2.4. Task 4

After merging the gene2go file and the geneIDs, the 3 tables of different gene scores with GO term ID, GO term Description and Go term count of each gene list are shown as table 6, 7 and 8 separately.

GO_ID	GO_term	GO_term_count
GO:0005515	protein binding	150
GO:0005634	nucleus	106
GO:0005654	nucleoplasm	84
GO:0005829	cytosol	66
GO:0005886	plasma membrane	58
GO:0005737	cytoplasm	55
GO:0006357	regulation of transcription by RNA polymerase II	45
GO:0046872	metal ion binding	42
GO:0045944	positive regulation of transcription by RNA polymerase II	38
GO:0000785	chromatin	38

Table 6: table of GO term counts with gene score 1

GO_ID	GO_term	GO_term_count
GO:0005515	protein binding	137
GO:0005886	plasma membrane	76
GO:0005634	nucleus	71
GO:0005829	cytosol	68
GO:0005737	cytoplasm	62
GO:0005654	nucleoplasm	56
GO:0016021	integral component of membrane	33
GO:0016020	membrane	32
GO:0005887	integral component of plasma membrane	31
GO:0070062	extracellular exosome	27

Table 7: table of GO term counts with gene score 2

GO_ID	GO_term	GO_term_count
GO:0005515	protein binding	333
GO:0005886	plasma membrane	179
GO:0005829	cytosol	141
GO:0005634	nucleus	126
GO:0005737	cytoplasm	122
GO:0005654	nucleoplasm	103
GO:0016021	integral component of membrane	89
GO:0005887	integral component of plasma membrane	82
GO:0016020	membrane	67
GO:0046872	metal ion binding	67

Table 8: table of GO term counts with gene score 3

3.2.5. Task 5

In this task, I obtained the bar chart data sets from the PantherDB tool[3] using the tab-delimited file with ID list as the first column, selected the organism Homo Sapiens, and then plotted the bar chart using seaborn package in python. The graphs that are exactly the same as the ones shown on the Panther website except with a descending order are as follow(Figure 4, 5 and 6):

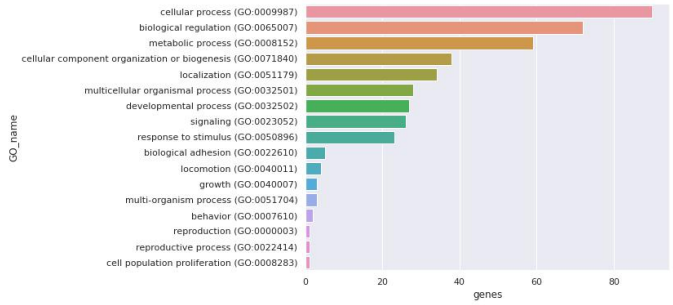

Figure 4: Hits of GO terms of genes with gene-score 1

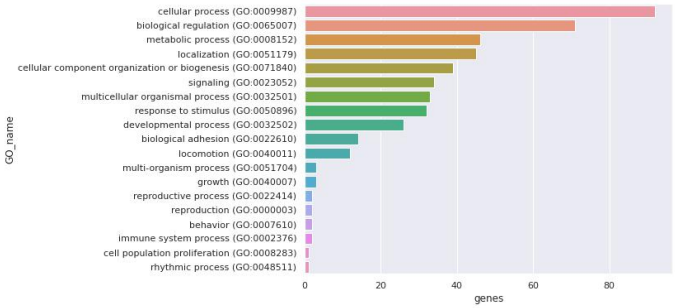

Figure 5: Hits of GO terms of genes with gene-score 2

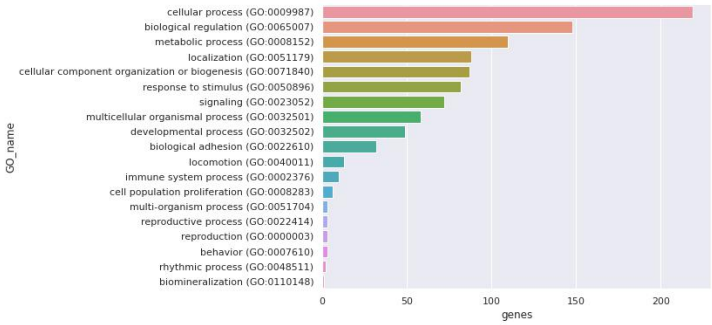

Figure 6: Hits of GO terms of genes with gene-score 3

Due to the page limitation, the figures may be a little unclear.

And I also plotted some other graphs showing and comparing the percentage of hits against total genes and total process hits(Figure 7, 8 and 9).

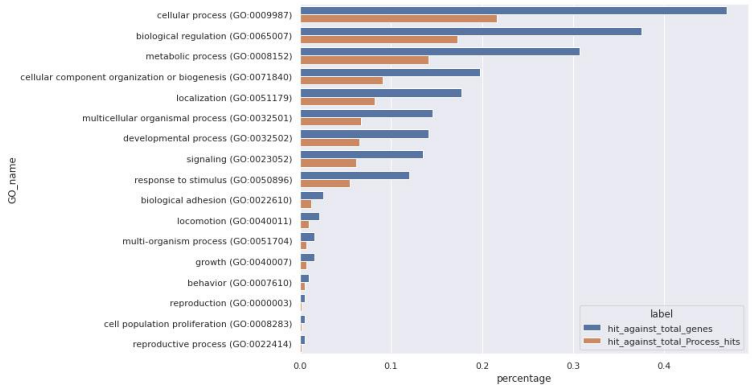

Figure 7: Percentage of hits of genes with gene-score 1

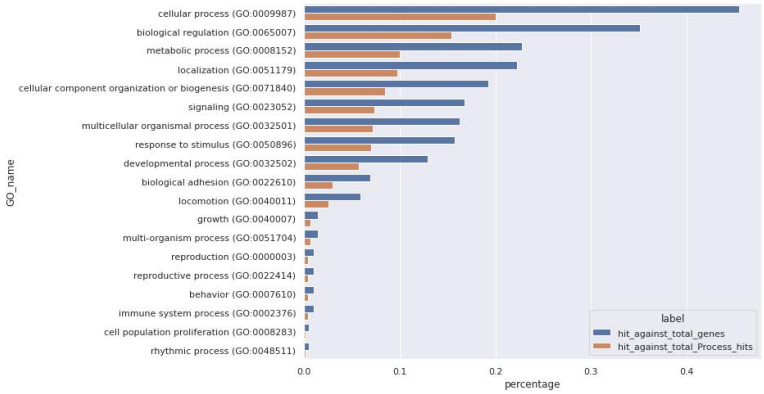

Figure 8: Percentage of hits of genes with gene-score 2

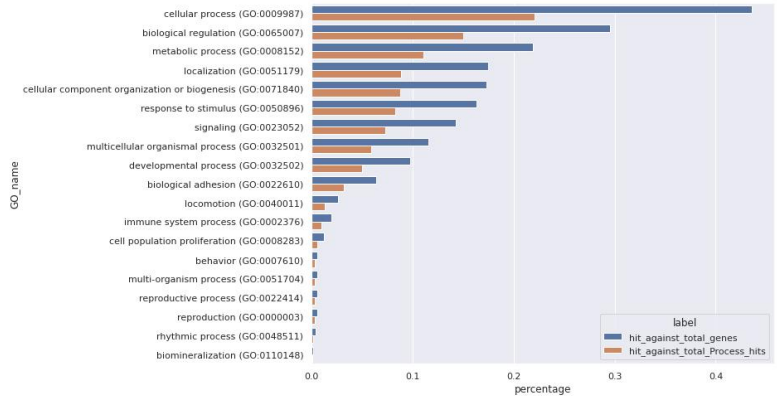

Figure 9: Percentage of hits of genes with gene-score 3

While submitting the gene lists in the website for the first time, the total genes shown on the website were always larger than the numbers of genes in the file. After some searching, I find that there are 2 possibilities that this situation happens:

1). **"**If for the biological process more genes are observed in the test list than expected, you have an over-representation of genes involved in induction of apoptosis. If fewer genes are observed than expected, you have an under-representation.**"**[7]

2). The ID list was in a wrong format. There are a limited number of IDs that are supported by the PantherDB Tools and they have to be in the right format.[8]

I did some tests and fond it clear that in this example, the problem is that the list I used was directly the UIDs of the genes while it was required to be *GeneID:XXX* where *XXX* represent the real gene ID.

The analysis of this part is presented in section 4.2.

3.3. Part 3 - Autism Gene Networks

3.3.1. Task 1

Using the STRING website, the statistics of gene-score 1 list are shown as below:

Number of nodes	Number of edges	Average node degree
193	990	10.3

Table 6: Statistics of gene score 2 list

3.3.2. Task 2

After downloading the tsv file and analyzed using python, the two biggest clusters are fond as cluster 1 and 2 both with 22 proteins.

Using the same way in part 2, the gene symbols in each cluster were mapped with the corresponding gene IDs, and the graphs plotted using the data from PantherDB were shown below:

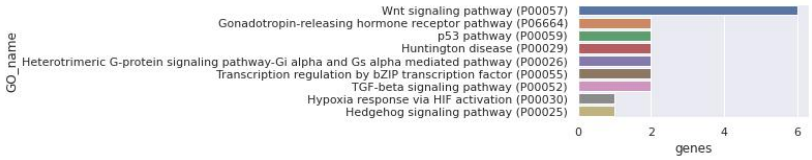

Figure 10: Hits of GO terms of cluster 1

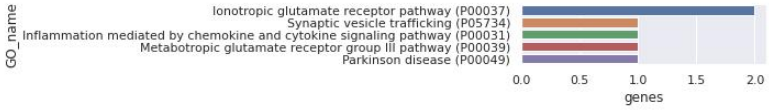

Figure 11: Hits of GO terms of cluster 2

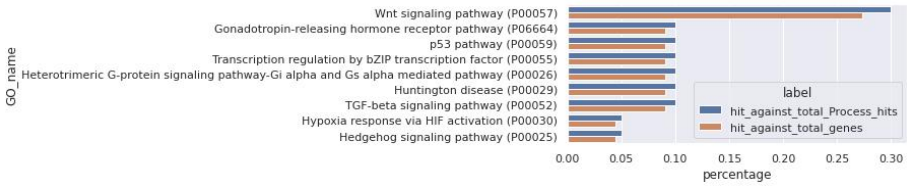

Figure 12: Percentage of hits of genes of cluster 1

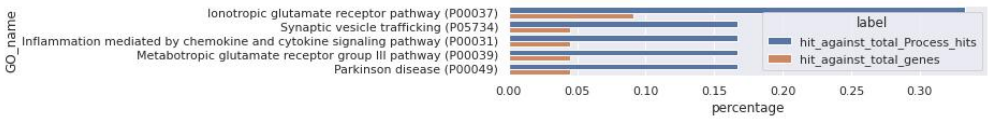

Figure 12: Percentage of hits of genes of cluster 2

3.3.3. Task 3

Since there is little space for the bitmap image, the graph is shown in the appendix A just to show the finish of the task.

The analysis of this part is presented in section 4.3.

3.3.4. Additional Analysis

After repeating the same process above for the other 2 gene score lists, it was fond that the number of genes in the biggest clusters of gene score 2(which are 14 and 9)

and 3(which are 15 and 12) are much less than that of gene score 1(22 and 22) which may refer to weaker connection between genes within one category.

The pathway GO terms hits of 3 gene score lists however, varies a lot while other ontologies are basically the same except with different hit rates.

4. Discussion

4.1. Part 1 - Autism Literature

Firstly, it is shown in 3.3.1 that there are almost the same numbers of genes in gene score categories 1 and 2 while the number of genes in category 3 is **2.5 times** greater. Secondly, from task 2 and 3, the top 5 genes with the most numbers of reports in the SFARI data set are **the same as** that from the PubMed search. It can be explained that these genes are the ones most likely to be related to Autism.

Among the top 5 genes, though it does not have the largest number of papers, *MECP2* appeared the earliest . But the number of papers related to this gene is not very significant since 2019. *SCN1A* also appeared early but the number of papers stayed low over the years and so did gene *SCN2A*. As for *NRXN1* and *SHANK3*, which are the top 2 SFARI genes, the number of papers increased from 2007 and are still very significant.

The genes are **fairly representative**. Since they include the genes that is related to the largest numbers of papers about autism which means they are most likely to be crucial to autism. And the time covered the most period of the history of the research about the relationship of autism and genes.[9]

However, it is still **not sufficient**. Firstly, there are still some genes with a relatively large numbers of papers for example, *PTEN* has 63 results which is just 4 papers less than *SCN1A*. Secondly, there are also many genes that is discovered to be related to autism since a very early time such as *PTEN* and *RELN*. In fact, from additional analysis(3.1.6), there are 10 genes that have papers related to autism before 2000 which is earlier than the most earliest paper of the top 5 genes, and there are even genes with papers in 1990s: *FMR1* which is far earlier than the earliest time these 5 genes covered. What's more, from additional analysis(3.1.6), there are still many genes beyond top 5 genes that have pepers related to autism in more recent years such as 2020 or 2019. The gene *CACNA1C* has though much less papers related to autism, the number of papers is increasing year by year which may be an evidence of its relationship to autism.

4.2. Part 2 - Autism Genes

Firstly, as was discussed in 3.2.1, there are genes that can not be mapped from the gene-info file using the gene symbols. From this task, it can be learned that, since each gene has only one unique UID, and many genes have aliases of pseudonyms, it is better to use UIDs instead of gene names and symbols.

Secondly, as one can tell from the results in 3.2.4 and 3.2.5, and also mentioned in the discuss forum, the GO terms in task 5 don't match the ones found in task 4. The ones in task 4 are much more specific whereas the pantherDB terms are quite broad. It is because that first, the pantherDB divided all the GO terms into some

different categories, and the one we chose was Biological Process. Second, those terms fond in task 5 are actually the parent terms of those in task 4 and if you click on the graph on the website, you will be able to get the children terms.

From task 4 and 5 in this part, the GO terms of each gene score are quit **similar** except with slightly different number of hit rates considering from 3.2.3 that the number of genes in gene score 3 list is 2.5 times as large as that of gene score 1 and 2. This means that these terms (**protein binding, nucleus, plasma membrane** from task 4 and **cellular process, biological regulation, metabolic process** from task 5) are very likely to be related to Autism.

The percentage of the gene hits against the total gene list is always about half the percentage of gene hits against the total process which means that the total hits were always about twice the numbers of genes.

Finally, the difficulty I encountered in task 5 also tells us that it is always a good habit to refer to the help documentations while using a tool that you are familiar with.

4.3. Part 3 - Autism Gene Networks

From task 1 in this part, we can see that there are 193 nodes(genes) and 990 edges which means that there are many **protein-protein interactions** between the SFARI genes(also from the data sets of gene score 2 and 3).

From the results in part 3, on the contrast to part 2, the 2 clusters have very different GO term hits. The first cluster is mainly about signaling and transcription and the second cluster is mainly about receptor. This shows that there are also several different clusters of genes in one gene score category. And Autism may be related not only to a special function but to a **combination of different genes**(which will affect many different proteins).

What's more, from the result in task 2 and the additional analysis, the difference between percentage of hits against total hits and the percentage of hits against total genes increases with the cluster number. This means that there are actually less hits in the larger clusters the reason of which is not discussed here.

From the additional analysis, it is also shown that the number of genes in the biggest clusters of gene score 2 and 3 are much less than that of gene score 1 which means that the **connections** between genes of score 1 are **closer** to that of genes of score 2 and 3. The pathway GO terms hits of 3 gene score lists however, varies a lot while the other ontologies are basically the same except with different hit rates.

5. References

[1] SFARI Gene, https://gene.sfari.org/

[2] NCBI Entrez Database, Homo_sapiens gene info, https://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

[3] NCBI gene2go file, https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz

[4] PantherDB tool, http://pantherdb.org/

[5] the STRING website, https://string-db.org/

[6] Medical Subject Heading, NIH, https://www.nlm.nih.gov/mesh/meshhome.html

[7] batch id supported ids help, Panther, http://pantherdb.org/tips/tips_batchIdSearch_supportedId.jsp

[8] Mi, Huaiyu et al., 2013. Large-scale gene function analysis with the PANTHER classification system. Nature protocols, 8(8), pp.1551–1566. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519453/

[9] Rylaarsdam, Lauren & Guemez-Gamboa, Alicia, 2019. Genetic Causes and Modifiers of Autism Spectrum Disorder. Frontiers in cellular neuroscience, 13, p.385.

6. Appendix A

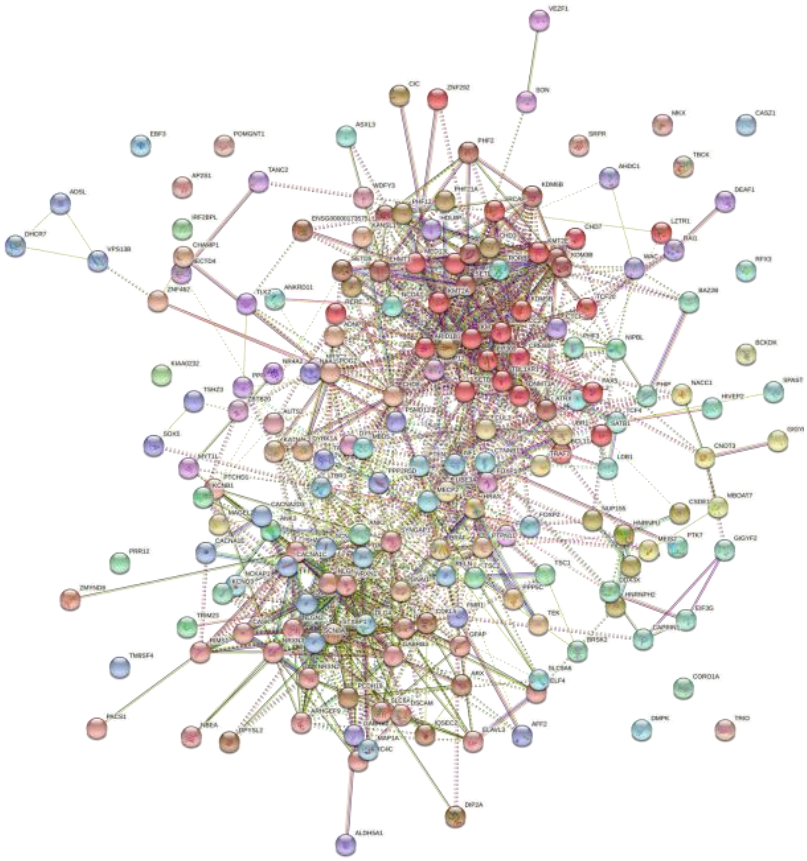

Figure 13: Bitmap image

Mark: all the code and files are in the git hub repository:
GIthub repository: https://github.com/Mushi0/Bioinformatic-2