Network Medicine@Data Science A.Y. 2020-2021

Epileptic drop attack

Simone ERCOLINO, Gabriele TAZZA, Aude MISSANA
Group no 7

Abstract

In this project, we are analyzing Epilepsy, a neural disease. It provokes sudden falls during which the patient is unable to control his body. It can have various causes: genetic, brain tumor, accidents and others.
Genetic researches have been conducted to identify the genes implicated, and get a better comprehension of this disease. Even if progress has been made through years, there are no results in the therapeutic field yet.
In this project, we will work with the genes involved in this disease. We select them according to their approved names, see the interactions among seed genes and with genes which have interactions with seed genes, and analyse them. Starting from the seed genes list linked to epilepsy (collected on DisGeNet) we conduct multiple analysis on the structure of their interactomes, in order to gain deeper knowledge on the disease and the interactions between the genes/proteins linked with it.

Basic introduction about the disease/process

Epileptic drop attack is a neural disease. It impacts around 50 millions persons in the world. It is one of the most frequent neural diseases in the world. It is due to an anormal activity of neural cells. These drop attacks are very disabling. The sudden falls, and convulsions can be very dangerous, for instance if the person is actually driving. In certain cases, the patients have to be accompanied all days long, or need a wheelchair. These seizures can last from seconds to minutes.
Actually, searchers have highlighted various causes for this disease. It can appear in relation to a genetic predisposition, after an important accident, as a consequence of brain tumor, and as a result of other pathologies. Dozens of causes exist, this disease can be diagnosed at each stage of a patient's life, and for both genders. In certain cases, for example with tumors, or malformations, it is possible to completely cure the patient. However, sometimes, this disease is not curable, for instance when it’s due to genetic origine.
Through years, researches have been able to isolate many genes impacted by this disease.

Seed genes

At the beginning, we were trying to get all the seed genes involved in the disease.
In the first hand, we had to get all the genes involved in the epileptic drop attack. We downloaded the datasets from DisGeNet website, and got a file containing the genes list for various diseases. We sorted it thanks to the disease ID : C0270846. We then put this list in the HGNC website to check if the name of each gene was approved. Some of them were “alias”, or “previous names”. We kept only the approved names for the rest of the project(file called “**hgnc-symbol-check_approved.csv**”).
With this new list of genes(file called “**uniprot_entry.txt**”), we got additional information on the seed genes (mainly protein name and geneID) from the dataset UniprotKB (collected at UniProt.Org).
Below an extract from the file “**uniprot_data_collection.csv**”.

Approved name	Entry	Protein names	GeneID
ABAT	P80404	4-aminobutyrate aminotransferase	18
ACAT1	P24752	Acetyl-CoA acetyltransferase	38
ACHE	P22303	Acetylcholinesterase	43
ADCYAP1	P18509	Pituitary adenylate cyclase-activating polypeptide	116
ADORA2A	P29274	Adenosine receptor A2a	135
ADRA1B	P35368	Alpha-1B adrenergic receptor	147

Table 1: Extract of the seed gene table (collected from UniprotKB

Summary on interaction data

In the website Biogrid human (collected at BioGRID), we started by downloading the information. After collecting them, we extracted only human/human interactions. To do this, we selected all the entries having both “Organism ID interactor A” and “Organism ID interactor B” equal to 9606 (code referred to the Homo Sapiens species). We put all of them into a table view for further operations(file called “**network_file.csv**”).
Once we had the list of the all human/human gene interactions from BioGRID, we checked if the component “Official Symbol Interactor A” or “Official Symbol Interactor B” corresponded with the ones in the list of seed genes from UNIPROT, in order to keep all the genes interacting with one seed gene. If they were, we kept them.
“Official Symbol Interactor A”, and “Official Symbol Interactor B” each represent a gene from an interaction. Once we have all of them, we have all the interactions.
We had to be careful and checked if we didn’t keep several similar interactions: in the table given by BioGRID, we can find an interaction between A and B, but also between B and A, in two different lines, even if they are the same. This problem is automatically solved by generating an undirected graph, which drops the doubled connections between the same genes.
The complete interactome of interest was stored in the file “**net_edge.csv**”, used to build the Graph object corresponding to our interactome.
We also created a file called “**interactome_genes.txt**” with all the genes found in the disease interactome.

N. seed genes collected in DisGeNet	101
N. seed genes found in Biogrid	100
N. interacting genes/proteins found	3923
N. interactions found (with double links)	183747

Table 2: Summary on seed genes and interactome data

Enrichment analysis (for the complete interactome)

We did the enrichment analysis with the website Enrichr to get the overrepresented GO categories and overrepresented pathways for the genes belonging to the complete disease interactome..
Below are reported the relative charts:

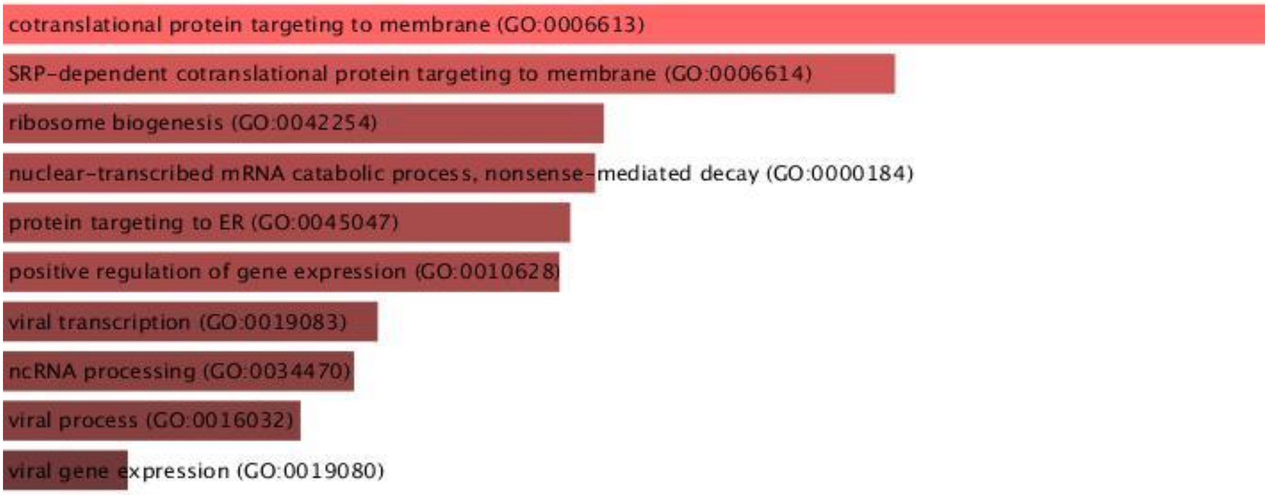

Figure 1: GO overrepresented biological process for all the genes in the disease interactome

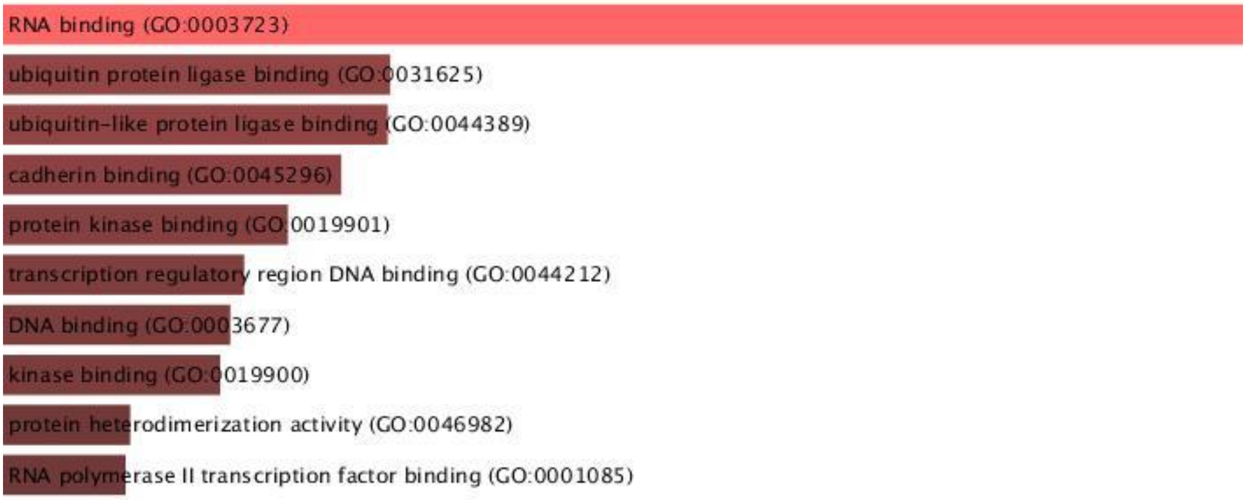

Figure 2: GO overrepresented molecular function for all the genes in the disease interactome

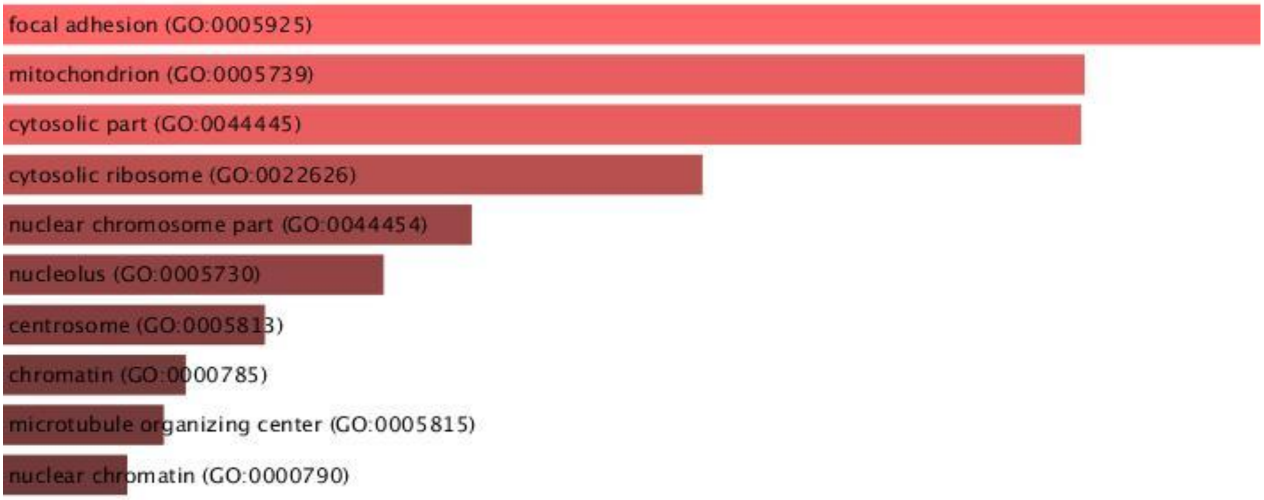

Figure 3: overrepresented GO cellular component for all the genes in the disease interactome

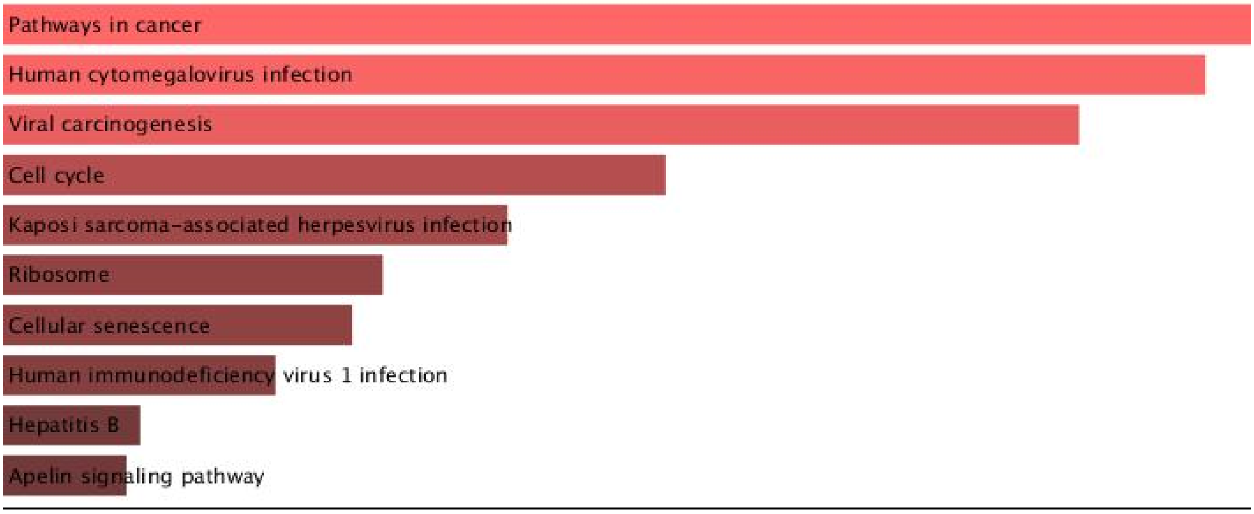

Figure 4: overrepresented pathways(KEGG 2019 human) for all the genes in the disease interactome

Interactomes data

To get the interactome graphs, we used the table we got previously which summarized the interactions. We started by creating a complete list of the edges of the interactions. We appended all the “Official Symbol Interactor A”, and “Official Symbol Interactor B” in a list of tuples.

Once we have this list, we use the module NetworkX to create a graph of the interactome, in order to measure centrality and other properties of genes in the interactome. We immediately discover that the disease interactome is a connected graph so its largest connected component (LCC) is the disease interactome itself.

Below we report the main information about this network, such as global measures and local measures for the first 20 nodes with the highest betweenness centrality(stored in file **“lcc_local.csv**”).

GLOBAL MEASURES	***LCC***
N. nodes	3923
N. links	130563
Average path length	2.4226
Average degree	66.5628
Average clustering coeff.	0.1857
Network diameter	6
Network radius	3
Centralization	0.3035

Table 3: global measures of LCC(largest connected component) of the disease interactome(in our case LCC=disease interactome).

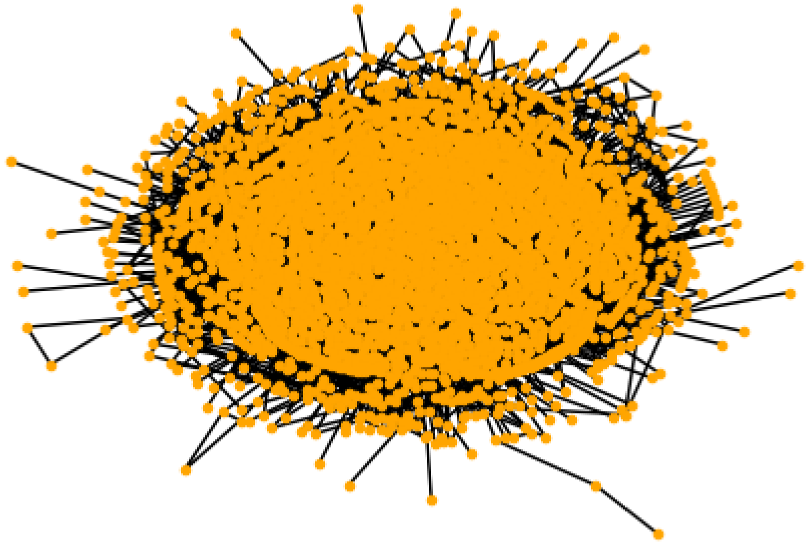

Figure 5: a figure of LCC(in our case all the disease interactome)

	betwenness centrality	node degree	eigenvector centrality	closeness centrality	betweenness /degree
APEX1	0.0874430332958	1016	0.04889768547635	0.559646118721	8,61E+10
PLEKHA4	0.03972978353041	1257	0.11858712997615	0.579149438865	3,16E+10
KIAA1429	0.03263073951851	988	0.09261909398070	0.559885795860	3,30E+10
APP	0.03056412633902	667	0.04882755622465	0.5357191640486	4,58E+10
TRIM25	0.02582904597993	819	0.08096774864096	0.5424619640387	3,15E+09
ESR2	0.02529966592884	898	0.07466900037601	0.5525500140884	2,82E+11
ELAVL1	0.02163002930508	657	0.0620380071765	0.5293561884194	3,29E+10
ESR1	0.02018787190692	930	0.09497870628732	0.5534857465424	2,17E+10
NTRK1	0.01985923410822	943	0.09692038368625	0.5568649723129	2,11E+11
HNRNPL	0.01857292334635	545	0.05147557515486	0.5242614623713	3,41E+11
MYC	0.01607264518672	928	0.10619451937347	0.5516950344633	1,73E+11
HNRNPH1	0.01310649846916	483	0.05298495832126	0.5204352441613	2,71E+11
TCF4	0.0127520885843	273	0.00990780899981	0.4616833431430	4,67E+10
KRAS	0.01270707476514	570	0.05232385740436	0.5176191104658	2,23E+10
KIF14	0.01240837776913	789	0.09180150616079	0.5344780594167	1,57E+10
EGFR	0.01237293407439	555	0.05422625494396	0.5249631910052	2,23E+09
RNF4	0.01121299395191	562	0.05410376391901	0.5230728194185	2,00E+11
TP53	0.01076922156214	607	0.07125206925441	0.5281443576622	1,77E+11
HIST1H4A	0.00961121549231	720	0.08773354083860	0.5298567954606	1,33E+11
XPO1	0.0092153675508	556	0.05696118567948	0.5152390961639	1,66E+11

Table 4: first 20 highest ranking genes of the LCC for betweenness centrality

Clustering methods for disease modules

We clustered the disease interactome with the MLC algorithm. To do it, we used the python module markov_clustering. Once it was done, we had to sort all the values to get only the ones we were interested in. We wanted the modules with the number of elements >= 10, and with a statistical overrepresentation of seed genes(hypergeometric test with a p_value < 0.05).

MODULE	N.of seed genes	N. total genes	N. seed genes/ N. total genes	p_value
Module 1	1	11	0.0909	0.0310
Module 2	1	10	0.1	0.0258
Module 3	1	10	0.1	0.0258
Module 4	1	11	0.0909	0.0310

Table 5: summary on disease modules emerged from clustering

Enrichment analysis on the disease modules

We did the enrichment analysis with the website Enrichr to get the overrepresented GO categories and overrepresented pathways for the genes belonging to each putative module. All the charts for this part are in the folder called "**clus_modules**"

DIAMOnD tool

The DIAMOnD tool script runs the DIAMOnD algorithm as described in **[1]**.
The DIAMOnD algorithm aims to identify the full disease module around a set of known disease proteins.
For the DIAMOnD tool, we used only the first 200 lines of the output as putative disease proteins for Epilepsy.
All file("**network_file.txt**": file with all the human/human interaction found in Biogrid, "**seed_file.txt**": file with all the 101 seed genes) used to run the DIAMOnD tool are stored in folder named "**DIAMOnD**". In this folder there is also the output file of DIAMOnD tool called "**first_200_added_nodes_weight_1.txt**" containing the first 200 genes ranked by DIAMOnd tool.
Below are reported the first 30 genes ranked by the DIAMOnD tool and the enrichment analysis conducted on the first 200 ranking (our putative disease module) in order to look for overrepresented GO categories and pathways.

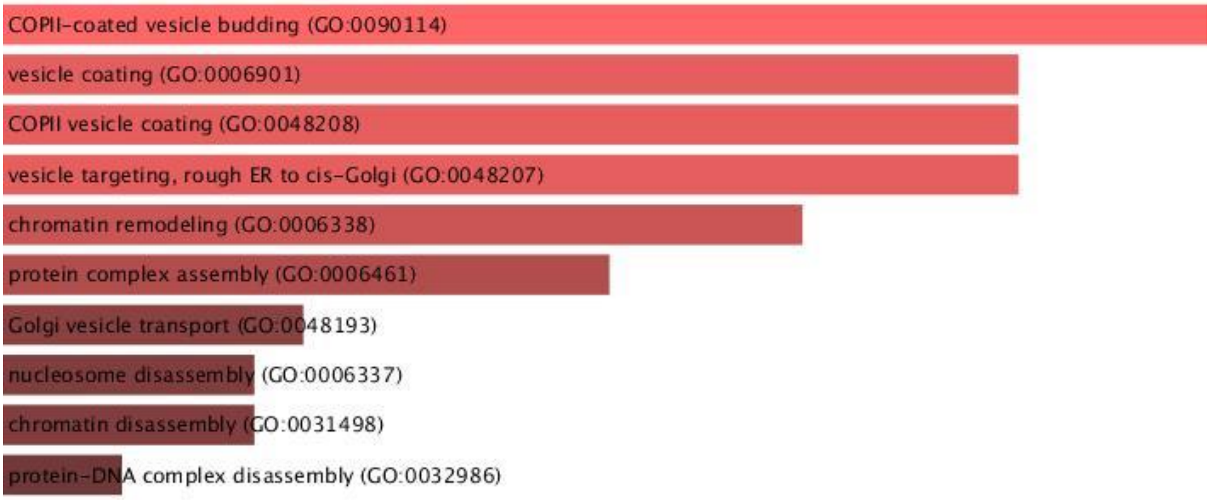

Figure 6: GO overrepresented biological process for DIAMOND putative disease module

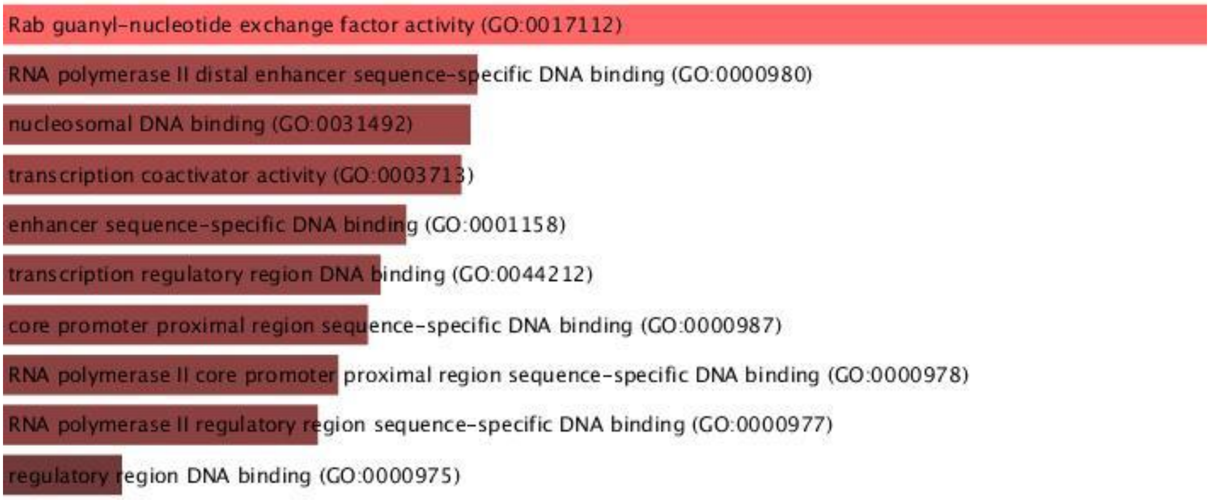

Figure 7: GO overrepresented molecular function for DIAMOND putative disease module

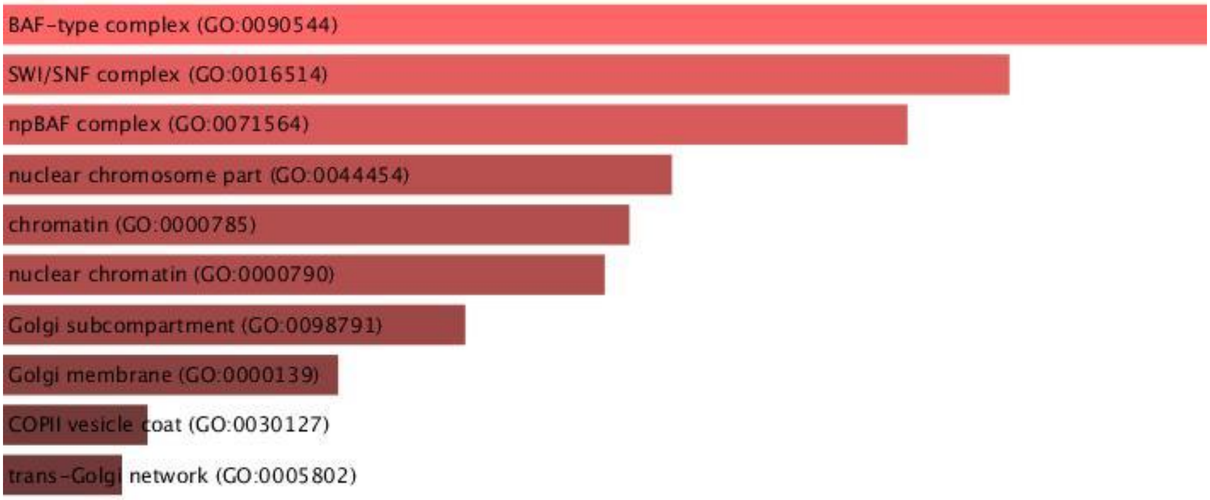

Figure 8: GO overrepresented cellular component for DIAMOND putative disease module

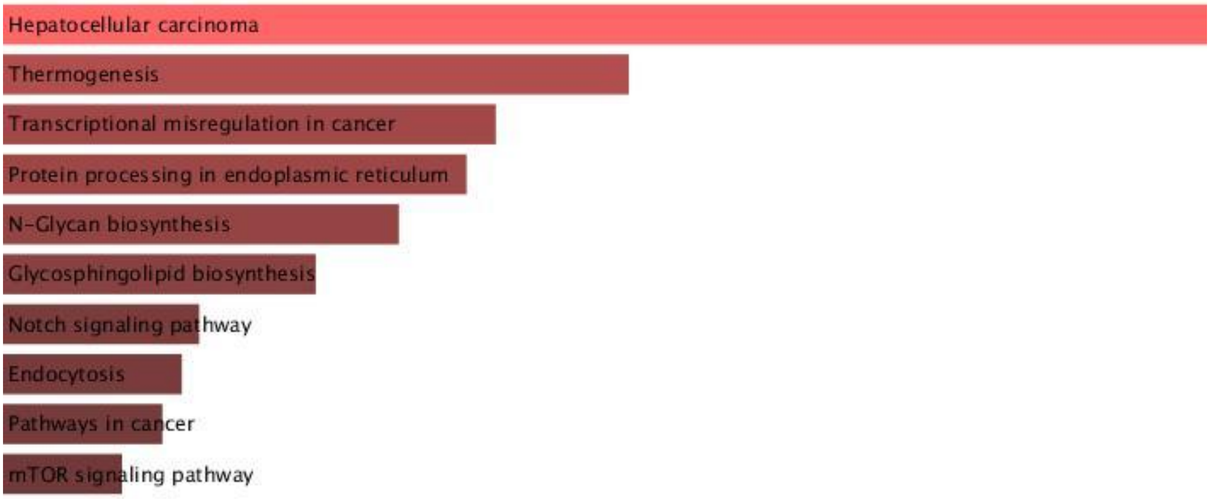

Figure 9: GO overrepresented pathways(KEGG 2019 human) for DIAMOND putative disease module

1	GPRASP1	7	GAS5	13	PAQR6	19	TRAPPC13	25	SEC24A
2	HRH3	8	WLS	14	KIAA1841	20	ARFGAP2	26	MAN2A1
3	HTR1D	9	OPRD	15	CHPT1	21	SLC39A7	27	SAR1A
4	VEGFA	10	GJA4	16	USF2	22	ARF1	28	SLC39A3
5	CLIC6	11	C9	17	TREM2	23	B4GALT1	29	TRAPPC2L
6	PDP1	12	TTL	18	SLC39A13	24	B4GALT3	30	MAN1A1

Table 6: List of the first 30 genes identified by the DIAMOnD tool

Notes and comments

During the project, we went across the gene PLPPR1. It was an “official name” so we kept it after checking it with the HGNC tool. This gene didn,’ have any function referenced in UNIPROT. However,
According to the website GeneCard: PLPPR1 Gene (Protein Coding) , this gene has several other aliases. One of them is PRG3, which is also an approved name. When checked, this one has a function in UNIPROT.
But when we compared its GeneID with the GeneID given by the information in the previous website, the 2 GeneID were different. Thus in the rest of the project, we chose to keep the gene PLPPR1. This gene disappeared from our list when we used the Biogrid website: indeed, PLPPR1 didn’t exist in the website.
The **code** used to perform all this tasks is called “**bio_project.ipynb**”

References

[1] (PDF) A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. PLoS Computational Biology, 2015.

references for the abstract and the introduction
Epilepsy - Symptoms and causes, Mayo Clinic, 2020
Drop Attack: Definition, Causes, Treatments, and More, Healthline, 2019