BIOINFORMATICS AND NETWORK MEDICINE

Final Project

Comparative assessment of disease gene prediction algorithms

Alessandro Quattrociocchi - 1609286

Tansel Simsek - 1942297

1. Abstract

In this report, we will interview different methodologies to investigate protein-protein interactions and its' role in the disease gene association. The aim is to find the most useful algorithm to extract correct genes to trace disease existence. We consider 5 different diseases, 5 different algorithms which are MCL (Markov Cluster), DIAMOnD (a disease module detection), DiaBLE, heat diffusion with Cytoscape bioinformatics software and Random Walk with Restart, and some of evaluation metrics such as precision, recall, F1 score and nDCG(normalized discounted cumulative gain). Also, we tried 6 additional diseases which all the results can be found in the additional pdf. As a final result, we investigated the most suitable disease as Autism Spectrum Disorder and the most powerful algorithm as DiaBLE with further enrichment analysis. All the related work can be found in https://github.com/BI-TeamProject/Final_project.

Keywords: GDA, PPI, MCL, DIAMOnD, DiaBLE, Cytoscape, Random Walk with Restart, Autism Spectrum Disorder.

2. Introduction

Protein-protein interaction has played a central role in bioinformatics in recent years, and the constant increase in computing power and machine learning techniques have facilitated an increasingly in-depth and large-scale study of interaction phenomena. In this paper, we present some results obtained from clustering and scoring algorithms applied to the humans' PPI represented as a graph. Once the interactome is created, we used the gene-disease associations as a ground-truth, to validate the clusters or the final ranking. The diseases considered are listed with the names and the respective MeSH disease class (multiple classes mean multiple disorders): *Autism Spectrum Disorders* (**F03**), *Intellectual Disability* (**C23;C10;F03;F01**), *Drug-Induced Liver Disease* (**C06;C25**), *Profound Mental Retardation(**C23;C10;F03;F01**)*, *Myocardial Failure* (**C14**). Here, we also want to introduce the reader to the structure of the article for a dive understanding of the presented models, the results, and the analysis. In addition to that, we want to point out some libraries used, and code structure created. The code is written in python 3.8 version by using class type of object to organize several functions. Networkx, pandas, numpy, statistics, and markov_clustering are some of the fundamental libraries utilized. Also, DIAMOnD algorithm implementation taken from the following *link*. Some of the main functions can be summarised as following, *preprocessing_dataset* filters the human interactions, removes the duplicates and self loops, *LCC_to_adj* extracts the largest connected component and creates the adjacency matrix of it, *MCL_evaluation_metrics* evaluates the performance metrics of the Markov Clustering algorithm, and *return_metrics* calculates the evaluation measures for the rest of the algorithm at different order.After this introductory section, there are sections such as the *material & methods* containing insights on how we pre-process the datasets and an in-depth explanation of the model's usage or implementation. Finally, the *results* section shows how we validated our findings, the enrichment analysis, further comments and interpretations on the analysis results.

3. Material and Methods

3.1 Pre-Processing

In this section, we report some fundamental steps to be followed for the reproducibility of the analysis.

As a first step, we downloaded the PPI interaction files from BioGRID and the disease genes association from DisGeNET. The PPI file is processed selecting all the genes related to the human genes only, removing the self-loops, and considering the nodes of the column 'Official Symbol Interactor A' as the source while those from 'Official Symbol Interactor B' as the target. Once having the data are ready from the processing, we move forward to the creation of the graph using the library **networkx** and, enumerating all the connected components of the graph, we selected the largest with **19618 nodes** and **665061 edges**. As a final step, we generated the graph's adjacency matrix, a $n \times n$ squared matrix which describes the topology of our directed graph where the element $a_{i,j}$ is 1 if there is an oriented path from node V_i to node V_j. Finally, the file containing the GDA (curated and extended) from DisGeNET doesn't need such extensive pre-processing as the one presented in the previous case, in fact, we simply filtered out all the genes from the GDA linked to the above-cited diseases.

3.1.1 GDA Overview

Disease name	*UMLS ID*	*MeSH class*	*# associated genes*	*# genes interactome*	*LCC size interactome*
Autism Spectrum Disorders	C1510586	F03	85	82	59
Intellectual Disability	C3714756	C23;C10;F03;F01	447	430	98
Drug-Induced Liver Disease	C0860207	C06;C25	404	319	89
Profound Mental Retardation	C0020796	C23;C10;F03;F01	139	134	71
Myocardial Failure	C1959583	C14	110	107	43

Table 1. Diseases overview.

3.2 MCL: Markov Clustering

Markov Clustering is a graph-based clustering algorithm, which has the advantage of not requiring in advance the knowledge of the number of clusters, like other clustering algorithms e.g. K-Means. The algorithm is based on the classic random walkers which, starting from a causal node, traverse the graph to reveal its topology. It may happen that a graph presents a high interconnection density between nodes that are sharing the same cluster, while a low-density value for nodes of different clusters. Iterating the random walk on the graph we want to determine what is the probability that starting from a node **u** we can reach a node **v**. Considering only one iteration, meaning that we are visiting each node of the graph at least once, the probability is given by $P_{uv} = \frac{1}{degree(v)}$. We define P as the matrix of the transition probabilities from each known U to a node V. The following procedure represents the core assumption of the Markov chain application in the algorithm: suppose to have a number of iterations > 1, to identify the nodes for which $P(k) << P(1)$ then bringing the probability of the step k to zero by raising it to a power greater than 1 and finally normalizing it. This procedure is called inflation and allows the transitions across the cluster to be highlighted. In this paper, we want to present an application of MCL for clustering human PPI. The MCL implementation takes as arguments the adjacency matrix of the graph, in our case the adjacency matrix from LCC and the value of inflation. To estimate the modularity (the higher the better), we tested all possible values in the range [1.5,2.7], obtaining that the best value of Q is **0.825425** when inflation is **1.8**. The whole tuning process took 7.31 hours by distributing the load over 3 different cores. The distribution of modularity is described in the following plot.

3.3 Cytoscape

Cytoscape is an open bioinformatics software for visualising molecular interaction networks and applying different algorithms to make several analysis. In this project we used it in order to implement diffusion based algorithm. The algorithm starts by selection of some nodes. Then, given time parameter the heat of the starting nodes will be distributed to the neighbors. Thus, there will be an output that shows diffusion output heat and ranking according to this number. Firstly, we did some preprocessing to the Biogrid_4.4.204.txt file that is already filtered by human. After utilising Connected Component Cluster application to reach Largest Connected Component, we confirmed that there are 19623 nodes and 896758 edges. For each gene and its' train datasets, we applied diffusion tool by taking t (time) parameter as 0.005. According to ranking output we calculated the metrics on validation datasets. It is important to note that we discarded the train dataset genes from the result so our first order rank belongs to the gene that took highest heat.

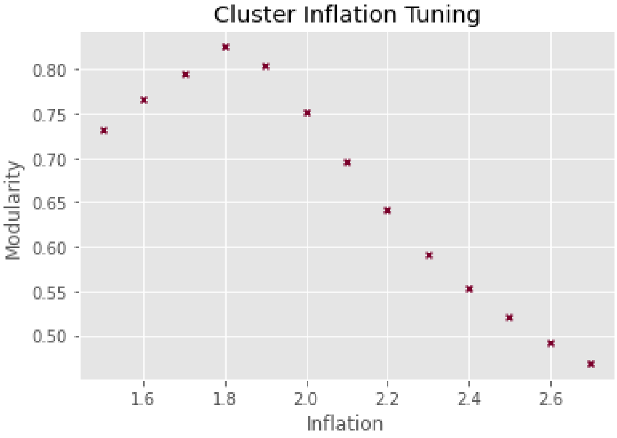


Figure 1. MCL modularity vs inflation

3.4 DIAMOnD: A DIseAse MOdule Detection

DIAMOnD is an algorithm that is acquired from a systematic analysis of connectivity patterns of disease proteins in Human Interactome by Susan Dina Ghiassian, Joerg Menche and Albert-Laszlo Barabasi. This algorithms considers the following parameters, the network(LCC), seed genes (disease genes), the maximum number of nodes to be added which will be the number of predicted genes at the end and alpha (given weight to the seeds). The algorithm starts by taking seed genes then at each step according to minimum p value the prediction results are getting extended. The p value calculation is made based on the Gaussian hyper-geometric function.

3.5 DiaBLE

DiaBLE is another version of the DIAMOnD algorithm. The difference can be observed as the following. While in DIAMOnD algorithm the universe is fixed and considering whole network, DiaBLE applies extended universe logic. DiaBLE starts with seed genes, its' candidate and candidate neighbor's. Then, at each step it makes a smaller expansion of the current seed set accordingly with the p-value. When the maximum number of nodes to be added parameter(N) gets larger, DiaBLE universe gets closer to the DIAMOnD universe. For the connected component networks, DiaBLE universe can be the same as DIAMOnD universe after some N.

3.6 RWR: Random Walk With Restart

Random walk with restart is the last algorithm that we present in this section. **RWR** exploits a classical iterative random walker on a graph that randomly samples the destination node starting from a given node. Here, some modifications to the classic model appear,specifically introduce a random walk with a restart probability of the walk in every step at node s with probability r, defined as following:

$$p^{t+1} = (1-r)Wp^t + rp^0 \tag{1}$$

where W is the L1 column normalizes adjacency matrix of the graph and p^t is a vector is the transition vector such that the i-th element holds the probability of being at node i at time t. In our case, p^0 was constructed using a uniform distribution that involves only the seed genes, in other works only those genes were selected as starting node with a probability equal to $\frac{1}{|seed\_genes|}$. Candidate genes were selected according to the steady-state probability vector p^∞ obtained when the difference between p^t and $p^t + 1$ (L1 norm) is $< 10^-6$.

4. Results

4.1 Evaluation Metrics pt.I: MCL

In this section, we present the techniques used to validate the algorithms presented above. Since we have adopted two different validations for MCL and the others (diamond, diable, Cytoscape, and Random Walk with Restart), they are reported separately. The first validation is that concerning MCL, developed after selecting the inflation equal to 1.8 as best inflation value which returned the higher modularity value among all those tested. Accordingly with this value, we created the clusters. To test whether a cluster is enriched or not, we divided the disease genes into 5 train-test different folds and, per each fold, if the selected cluster share at least 2 seed genes with the current train fold, we estimated the probability value of the extraction (without re-introduction) of a seed gene with an hyper-geometric distribution, as follows:

$$p_x(X) = Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{2}$$

If $p < 0.05$ the cluster is considered enriched and the following metrics are calculated:

- **TP**: number of probe set genes in all enriched clusters
- **FP**: number of genes in all enriched cluster which are not seed gene
- **FN**: number of probe seed genes not present in any enriched clusters

As this type of algorithm only returns genes contained in the cluster without any ranking, the metrics considered are accuracy, recall and F1 score. Below are the results obtained with MCL for the five diseases considered. Please note that we have extended the validation by considering more disease genes in the pdf with the overall results.

Disease	*Precision*	*Recall*	*F1 score*
Autism Spectrum Disorders	1.23	0.13	0.24
Intellectual Disability	7.17	0.77	1.39
Drug-Induced Liver Disease	7.55	0.81	1.46
Profound Mental Retardation	2.42	0.26	0.47
Myocardial Failure	0.33	0.04	0.06

Table 2. Final Results overview.

4.2 Evaluation Metrics pt.II: DIAMOnD, DiaBLE, Cytoscape, Random Walk with Restart

Since DIAMOnD, DiaBLE, Cytoscape and Random Walk with Restart algorithms return the predictions as a ranking, evaluation metrics for these are computed in the common way. In order the compute the metrics, there are some building blocks firstly measured before evaluation. These are considered by taking first k positions, and details can be the following:

- **TP @ k**: number of probe genes in the first k positions of the ranking
- **FP @ k**: number of genes that are in the top k positions but not probe genes
- **FN @ k**: number of probe seed genes not present in the top k positions
- **k** : 50, n/10, n/4, n/2, n where n : number of disease's GDAs

The calculations are made for each folds, which we applied 5-fold cross validation at the beginning to the probe genes(curated gene-disease associations for the first part, and all_gene_disease_associations.tsv file for the extended validation part). In the DIAMOnD and DiaBLE train folds taken as seed_genes parameter, in the Cytoscape train folds selected to apply heat diffusion, in the Random Walk with Restart train folds given as disease genes input. Then, after getting prediction as a rank and finding the above rates, we executed Precision, Recall and F1 scores by considering the corresponding test folds.

Moreover, we considered normalized Discounted Cumulative Gain(nDCG) at k as a evaluation metric as well where the formulation can be given as following:

$$\textbf{DCG @ k} = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)}, \tag{3}$$

$$\textbf{iDCG @ k} = \sum_{i=1}^{|REL_k|} \frac{rel_i}{log_2(i+1)}, \tag{4}$$

$$\textbf{nDCG @ k} = \frac{DCG@k}{iDCG@k}, \tag{5}$$

$rel_i = 1$ if i^{th} prediction is a probe gene, $0 =$ otherwise,
$REL_k =$ list of probe genes in the ground truth up to position k.

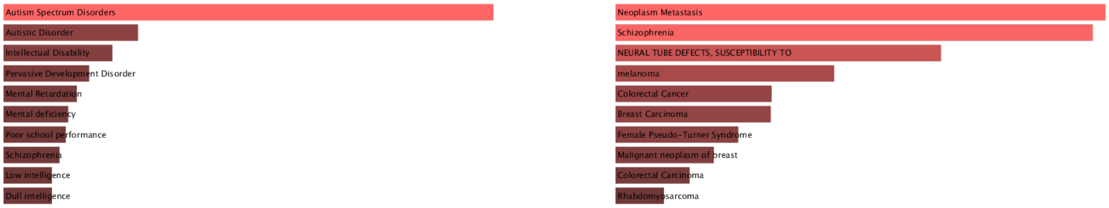
4.3 Final Resul Table

Autism Spectrum Disorders

@	Eval Metric	Diamond	Diable	Cytoscape	RWR	Diamond Ex	Diable Ex	CytoscapeEx	RWR Ex
50	P	1.6 ± 1.67	2.0 ± 1.41	0.0 ± 0.0	0.0 ± 0.0	5.2 ± 3.35	5.2 ± 3.35	4.4 ± 8.76	18.0 ± 8.25
	R	4.71 ± 4.92	5.88 ± 4.16	0.0 ± 0.0	0.0 ± 0.0	1.21 ± 0.78	1.21 ± 0.78	1.03 ± 2.05	4.2 ± 1.93
	F1	3.98 ± 1.72	3.73 ± 1.49	0.0 ± 0.0	0.0 ± 0.0	2.46 ± 0.73	2.46 ± 0.73	4.17 ± 4.82	6.81 ± 3.12
	nDCG	1.96 ± 2.02	2.45 ± 1.7	0.0 ± 0.0	0.0 ± 0.0	6.53 ± 4.91	6.53 ± 4.91	4.2 ± 8.39	17.39 ± 7.62
n/10	P	5.0 ± 6.85	7.5 ± 6.85	0.0 ± 0.0	0.0 ± 0.0	2.8 ± 1.75	2.8 ± 1.75	2.62 ± 4.35	12.52 ± 4.21
	R	2.35 ± 3.22	3.53 ± 3.22	0.0 ± 0.0	0.0 ± 0.0	1.4 ± 0.87	1.4 ± 0.87	1.31 ± 2.18	6.26 ± 2.11
	F1	8.0 ± 0.0	8.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	2.33 ± 0.6	2.33 ± 0.6	2.91 ± 3.43	8.34 ± 2.81
	nDCG	4.33 ± 6.07	5.93 ± 5.68	0.0 ± 0.0	0.0 ± 0.0	4.12 ± 2.93	4.12 ± 2.93	2.91 ± 5.1	13.5 ± 4.4
n/4	P	1.9 ± 2.61	2.86 ± 2.61	0.0 ± 0.0	0.0 ± 0.0	1.87 ± 0.75	1.87 ± 0.75	2.62 ± 3.01	6.29 ± 1.66
	R	2.35 ± 3.22	3.53 ± 3.22	0.0 ± 0.0	0.0 ± 0.0	2.34 ± 0.94	2.34 ± 0.94	3.27 ± 3.75	7.84 ± 2.08
	F1	5.26 ± 0.0	5.26 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	2.08 ± 0.83	2.08 ± 0.83	2.91 ± 3.34	6.98 ± 1.85
	nDCG	2.36 ± 3.3	3.23 ± 3.09	0.0 ± 0.0	0.0 ± 0.0	2.73 ± 1.46	2.73 ± 1.46	2.77 ± 3.63	7.89 ± 2.07
n/2	P	1.9 ± 1.99	2.38 ± 1.68	0.0 ± 0.0	0.0 ± 0.0	1.79 ± 0.92	1.79 ± 0.92	2.06 ± 1.52	4.15 ± 1.02
	R	4.71 ± 4.92	5.88 ± 4.16	0.0 ± 0.0	0.0 ± 0.0	4.48 ± 2.31	4.48 ± 2.31	5.14 ± 3.8	10.37 ± 2.55
	F1	4.52 ± 1.96	4.24 ± 1.69	0.0 ± 0.0	0.0 ± 0.0	2.56 ± 1.32	2.56 ± 1.32	2.94 ± 2.17	5.93 ± 1.46
	nDCG	2.2 ± 2.28	2.75 ± 1.92	0.0 ± 0.0	0.0 ± 0.0	2.3 ± 1.27	2.3 ± 1.27	2.23 ± 2.12	5.43 ± 1.33
n	P	0.94 ± 0.98	1.18 ± 0.83	0.47 ± 0.64	0.47 ± 0.64	1.4 ± 0.52	1.4 ± 0.52	1.7 ± 0.96	2.91 ± 0.45
	R	4.71 ± 4.92	5.88 ± 4.16	2.35 ± 3.22	2.35 ± 3.22	7.0 ± 2.61	7.0 ± 2.61	8.5 ± 4.82	14.57 ± 2.28
	F1	2.61 ± 1.13	2.45 ± 0.98	1.96 ± 0.0	1.96 ± 0.0	2.34 ± 0.87	2.34 ± 0.87	2.83 ± 1.6	4.86 ± 0.76
	nDCG	1.35 ± 1.4	1.69 ± 1.18	0.36 ± 0.49	0.35 ± 0.47	1.75 ± 0.65	1.75 ± 0.65	1.85 ± 1.39	3.82 ± 0.75

5. Enrichment Analysis

Enrichment analysis is the last step in our discussion. Once the process of validating the results of each algorithm had been completed, we selected the disease-algorithm pair with the best results and feed the algorithm with the entire set of disease genes available. Autism Spectrum Disorder is a disease to be analyzed, as it is one of the top performance diseases and great interest to us for further study. The number of genes belonging to autism disorder is 85 of which 82 are part of the LCC. The DiaBLE is the algorithm selected, imposing the number of predictions equal to 200, resulting in a disease module (disease genes + predicted disease genes). We enriched the genes using Enrichr by pasting the diseases into the appropriate window, moving to the Disease/Drugs section, and selecting the DisGeNET enrichment. To enhance the differences between the seed genes and the predicted putative genes, we enriched the both sets. The result of the enrichment analysis is shown below.



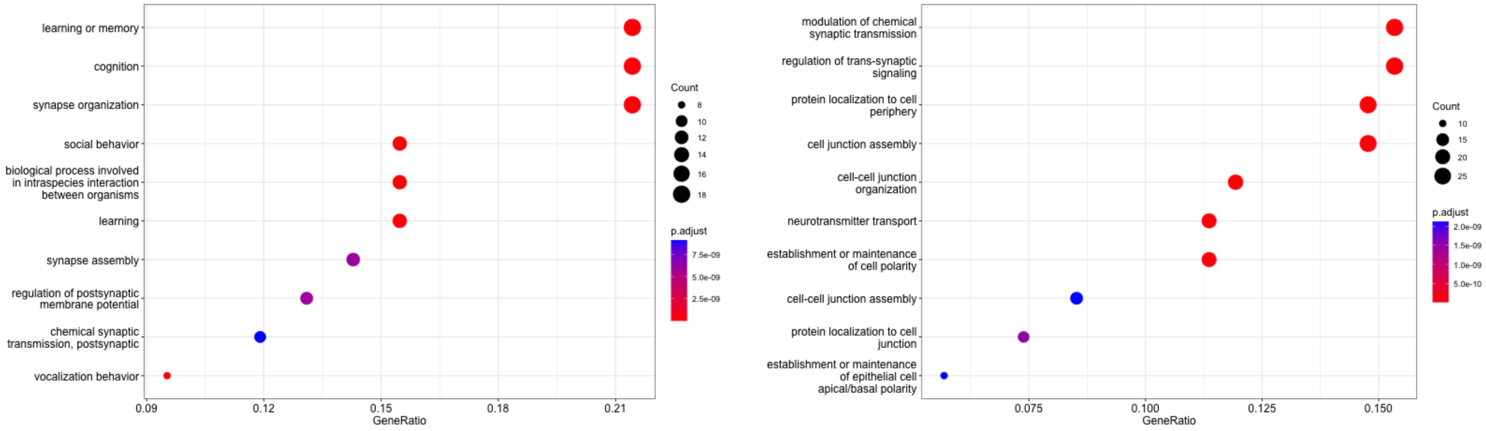
(a) seed genes

(b) putative genes

Figure 2. Enrichr Diseases/Drugs Enrichmenten

The enrichment of the seed genes shows in the first position the autism spectrum disorder, as we expected but other disorders as well, such as: intellectual disability, pervasive development disorder, mental retardation and schizophrenia. At the top position of the putative disease genes enrichment we find: Neoplasm Metastatis, Schizophrenia, Neural tube defects. Albeit the two previous plots may show several diseases sharing a set of predicted genes, we extended the enrichment by considering GO ontologies from Bioconductor. The R-code can be found in the additional material folder on GitHub. Here the differences taper off and similar pathways can be identified, such as those regulating postsynaptic transmission, maintenance of cell polarity, transport of neurotransmitters, ion channel activity and ion channel complex. Those findings are enhanced by the adjusted p-value that is of the order of 10^{-9} or 10^{-5}, showing statistical significance. As a further validation we provide some references. In [1] authors showed how regulation of membrane potential may appear in the process of mitochondrial respiratory capacity and membrane potential, reporting a significantly unregulated energy metabolism pathway in a lymphoblastoid cell line. In [2] is showed how mutated genes directly or indirectly affect the postsynaptic neuronal compartment. Finally, [3] explain how PDZ domain is involved multiple biological processes such as transport, ion channel signaling, and other signal transduction systems, that have been studied as potential ASD-related pathways in recent research fields. Finally, in the final positions of MF, ”glutamate receptor binding” is reported which, although it has a very high adjusted p-value, is of considerable importance in the study of autism spectrum disorder, as discussed in [4] .

5.1 Go Enrichment: Biological Process

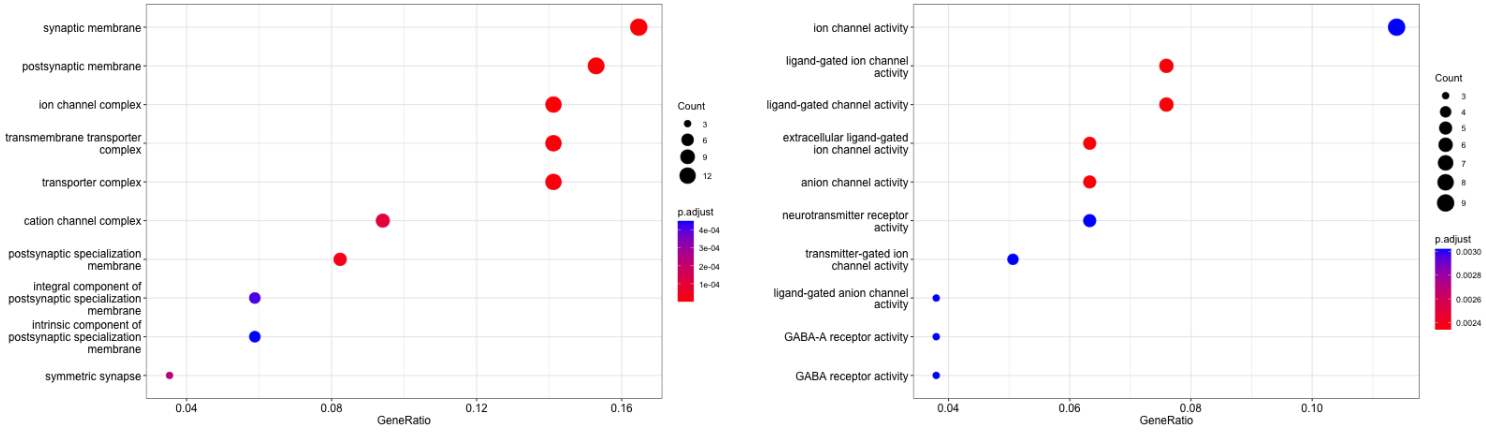


(a) seed genes

(b) putative genes

Figure 3. BP ClusterProfiler GO Enrichment Analysis

5.2 CC: Cellular Component

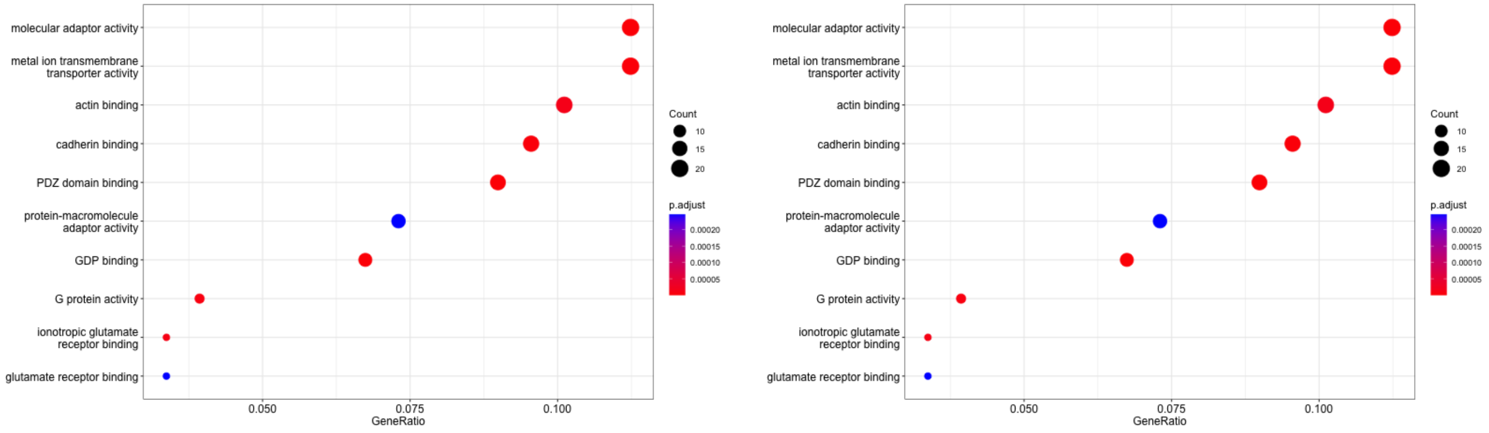


(a) seed genes

(b) putative genes

Figure 4. CC ClusterProfiler GO Enrichment Analysis

5.3 MF: Molecular Function



(a) seed genes

(b) putative genes

Figure 5. MF ClusterProfiler GO Enrichment Analysis

6. Further Materials

We extended the validation to a total of 11 genes, all tables of the results are attached to the article, or downloadable here. For the full code and raw data, please check https://github.com/BI-TeamProject/Final_project.

References

[1] Hassan H, Zakaria F, Makpol S, Karim NA. A Link between Mitochondrial Dysregulation and Idiopathic Autism Spectrum Disorder (ASD): Alterations in Mitochondrial Respiratory Capacity and Membrane Potential. Curr Issues Mol Biol. 2021 Dec 16;43(3):2238-2252. doi: 10.3390/cimb43030157. PMID: 34940131.

[2] Paola Bonsi, Antonella De Jaco, Laurent Fasano, Paolo Gubellini, Postsynaptic autism spectrum disorder genes and synaptic dysfunction, Neurobiology of Disease, Volume 162, 2022, 105564, ISSN 0969-9961

[3] Lee, HJ., Zheng, J.J. PDZ domains and their binding partners: structure, specificity, and modification. Cell Commun Signal 8, 8 (2010). https://doi.org/10.1186/1478-811X-8-8

[4] Rojas DC. The role of glutamate and its receptors in autism and the use of glutamate receptor antagonists in treatment. J Neural Transm (Vienna). 2014;121(8):891-905. doi:10.1007/s00702-014-1216-0

[5] A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome Ghiassian SD, Menche J, Barabási AL (2015). PLOS Computational Biology 11(4): e1004120. https://doi.org/10.1371/journal.pcbi.1004120

[6] A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome Ghiassian SD, Menche J, Barabási AL (2015). PLOS Computational Biology 11(4): e1004120. https://doi.org/10.1371/journal.pcbi.1004120

[7] Walking the Interactome for Prioritization of Candidate Disease Genes Sebastian Köhler, Sebastian Bauer, Denise Horn, Peter N. Robinson Open ArchivePublished:March 28, 2008DOI:https://doi.org/10.1016/j.ajhg.2008.02.013

[8] Petti M, Bizzarri D, Verrienti A, Falcone R, Farina L. Connectivity Significance for Disease Gene Prioritization in an Expanding Universe. IEEE/ACM Trans Comput Biol Bioinform. 2020 Nov-Dec;17(6):2155-2161. doi: 10.1109/TCBB.2019.2938512. Epub 2020 Dec 8. PMID: 31484130.

[9] Markov Clustering Git Folder: GuyAllard/markov_clustering

[10] DIAMonD Git Folder: dinaghiassian/DIAMOnD