Putative Disease Gene Identification and Drug Repurposing for Hypertensive Disease

Boesso Simone, Lavagna Leonardo, Skerl Leonardo

Group 4

Abstract

The aim of this study is to find new drugs, as well as currently used drugs, associated to this disease using a network approach. We used different algorithms: DIAMOnD, DIaBLE and Diffusion Based. We evaluated each algorithm using standard cross-validation techniques and we found that the latter perform better. We used it to find the 200 best putative genes, carried out enrichment analysis with different tools, and from them we selected the best 20 to find the highest matching drugs for each such gene. We found 6 drugs with the same frequencies, two of them were already discussed in several studies in relation with Hypertensive Disease while the other four weren't. We also tried to find new putative disease modules from the human interactome network and we found eight list of possible disease associated genes.

1 INTRODUCTION

The Hypertensive disease is a high blood pressure condition that affects the body's arteries. Blood pressure is determined by two things: the amount of blood the heart pumps and how hard it is for the blood to move through the arteries. There are two main types of high blood pressure (cfr. [4]):

- Primary hypertension: if there's no identifiable cause of high blood pressure
- Secondary hypertension: This type of high blood pressure is caused by an underlying condition. It tends to appear suddenly and cause higher blood pressure than does primary hypertension.

We have different risk factor for hypertension: Age, Race, Family History, Obesity, Lack of Exercise, Smoke, Salty Food, Alcohol, Stress, Pregnancy. The last one is a real problem affecting like the 3% of women. To treat this disease, in terms of medical treatments, they use different types of medicine like Alpha Blockers, Beta Blockers, Aldosterone antagonists, Renin inhibitors and Vasodilators. The aim of this research is to show that with a network approach, without relying on medical data, we can find drugs already used to treat Hypertensive Disease in medical practice.

2 MATERIALS AND METHODS

For the following tasks we decided to use ”Python 3.0” as programming language in conjunction with google colab and google drive. We used standard libraries for data processing and analysis (e.g. `pandas`, `statistics`, `numpy`). All our network based studies were performed with `networkx` package. We used also custom code developed from scratch and available from the package auxiliary functions which was created by our group. Other important algorithms we used are DIAMOnD [5] and DIaBLE, both taken from ”Github” repositories [7]. For the Enrichment analysis we used EnrichR, while for the diffusion based algorithm we used Cytoscape [9]. All the project's code and files can be found in [8]

2.1 PPI and GDA data gathering and interactome reconstruction

In order to obtain the list of drugs we downloaded the Protein-Protein-Interaction dataset (PPI) from BioGRID [10] and we extracted all the”Homo Sapiens” genes, keeping only the physical interactions and dropping all redundancies (e.g. self-loops). After this preprocessing, we obtained the human interactome network (HSN). To obtain the disease network we download the disease-genes association (DGA) from the 2 networks DGA and HSN we extracted the genes in the disease PPI (DGinPPI) and we created the seed gene's list. From DGinPPI we constructed the disease network (DN) as follows:

- the nodes are the elements in DGinPPI
- the edges are taken from the intersection between DGA and HSN

At this point, we computed the disease LCC of DN and that is the LCC. With our two datasets we proceeded to check the presence of the genes in our PPI network obtaining the results in Figure 1. The principal metrics of our disease LCC network can be found in Figure 2.

2.2 Putative disease genes identification algorithms

The next phase of the research is to predict with a network approach all Putative disease genes. In order to do that we tried different algorithms:

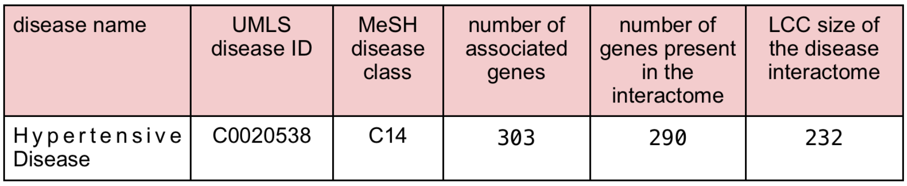
- DIAMOnD;
- DIaBLE;
- Diffusion Based Algorithm with Cytoscape;

We evaluated each model according to the following steps:

- We split at random the seed gene list into 5 parts
- We performed a 5-Fold Cross Validation
- At each cross we used one split as the probe set and the other four as the training set

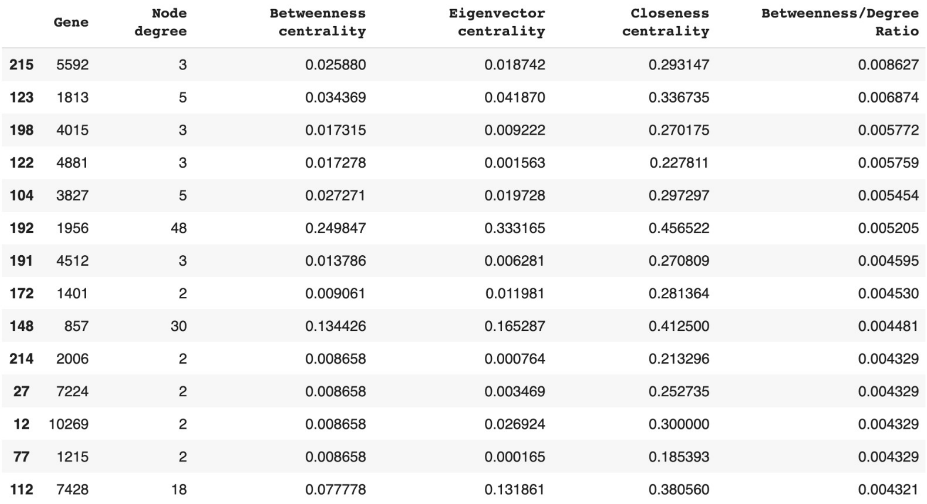
- We computed the mean over all the crosses of the following metrics:
  - Precision
  - Recall
  - F1-Score

In the next section we'll see the results of these evaluations.



disease name	UMLS disease ID	MeSH disease class	number of associated genes	number of genes present in the interactome	LCC size of the disease interactome
Hypertensive Disease	C0020538	C14	303	290	232

Figure 1: Summary of GDAs and basic network data



	Gene	Node degree	Betweenness centrality	Eigenvector centrality	Closeness centrality	Betweenness/Degree Ratio
215	5592	3	0.025880	0.018742	0.293147	0.008627
123	1813	5	0.034369	0.041870	0.336735	0.006874
198	4015	3	0.017315	0.009222	0.270175	0.005772
122	4881	3	0.017278	0.001563	0.227811	0.005759
104	3827	5	0.027271	0.019728	0.297297	0.005454
192	1956	48	0.249847	0.333165	0.456522	0.005205
191	4512	3	0.013786	0.006281	0.270809	0.004595
172	1401	2	0.009061	0.011981	0.281364	0.004530
148	857	30	0.134426	0.165287	0.412500	0.004481
214	2006	2	0.008658	0.000764	0.213296	0.004329
27	7224	2	0.008658	0.003469	0.252735	0.004329
12	10269	2	0.008658	0.026924	0.300000	0.004329
77	1215	2	0.008658	0.000165	0.185393	0.004329
112	7428	18	0.077778	0.131861	0.380560	0.004321

Figure 2: Summary of GDAs and basic network data

2.3 Putative disease genes identification

In this part we're interested in using some community detection algorithms in order to identify putative disease modules. We used Markov Clustering (MCL [6]) with the default inflation parameter and applied it to the human interactome LCC. Due to computational limitations (e.g. limited RAM) we couldn't try different inflation values. Once the communities were found we carried out the following steps:

- We considered only the communities with a number of nodes between 10 and 1000
- We used an Hypergeometric test if it is a putative disease module or not
- We listed the communities that we found out as putative disease module

2.4 Best algorithm choice and putative disease gene identification

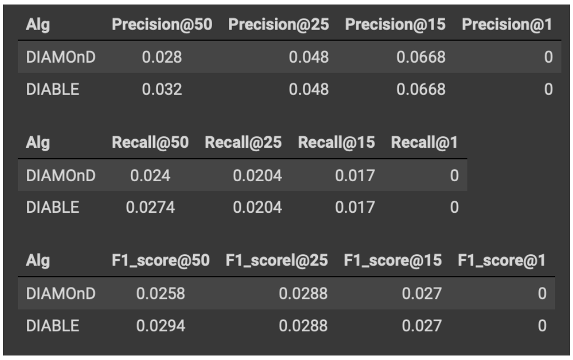
After our analysis in the previous point we found out that Diffusion Based Algorithm (for t = 0.01) is the one that performs better. We used it to predict new putative disease genes using all known GDAs as seed genes and we obtained a list of predicted genes. Over this list we selected the first 200 putative disease genes and proceeded with the Enrichment Analysis using EnrichR platform on GO-BP, GO-MF, GO-CC, KEGG-Pathways.

2.5 Drug repurposing

In this section we're interested in associating genes to drugs. So we focused on the first 20 putative disease genes in the ranking obtained in the previous point. We inserted the genes in the DGIdb [11] table in order to discover the associated diseases to each gene. In this step we faced a nomenclature problem: of the 20 genes we searched just 18 were founded. ACT wasn't in DGIdb database so we've tried to search for synonyms of the gene, but it seems there isn't any. We decided to not include it instead of using other genes with similar names (like ACTA1). TRNV wasn't also in the DGIdb database but we found a synonym: MT-TV. In the next part we'll discuss the results we obtained.

3 RESULTS AND DISCUSSION

In this section we can see the results of our study. For first let's take a look to the performances of each algorithm:

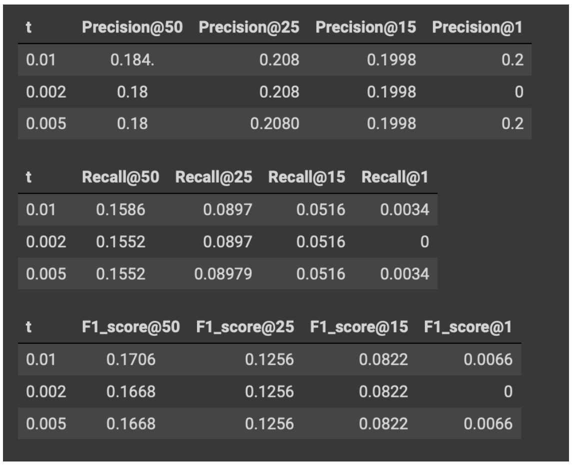


Alg	Precision@50	Precision@25	Precision@15	Precision@1
DIAMOnD	0.028	0.048	0.0668	0
DIABLE	0.032	0.048	0.0668	0

Alg	Recall@50	Recall@25	Recall@15	Recall@1
DIAMOnD	0.024	0.0204	0.017	0
DIABLE	0.0274	0.0204	0.017	0

Alg	F1_score@50	F1_scorel@25	F1_score@15	F1_score@1
DIAMOnD	0.0258	0.0288	0.027	0
DIABLE	0.0294	0.0288	0.027	0

Figure 3: Performances for Diamond and DIaBLE



t	Precision@50	Precision@25	Precision@15	Precision@1
0.01	0.184.	0.208	0.1998	0.2
0.002	0.18	0.208	0.1998	0
0.005	0.18	0.2080	0.1998	0.2

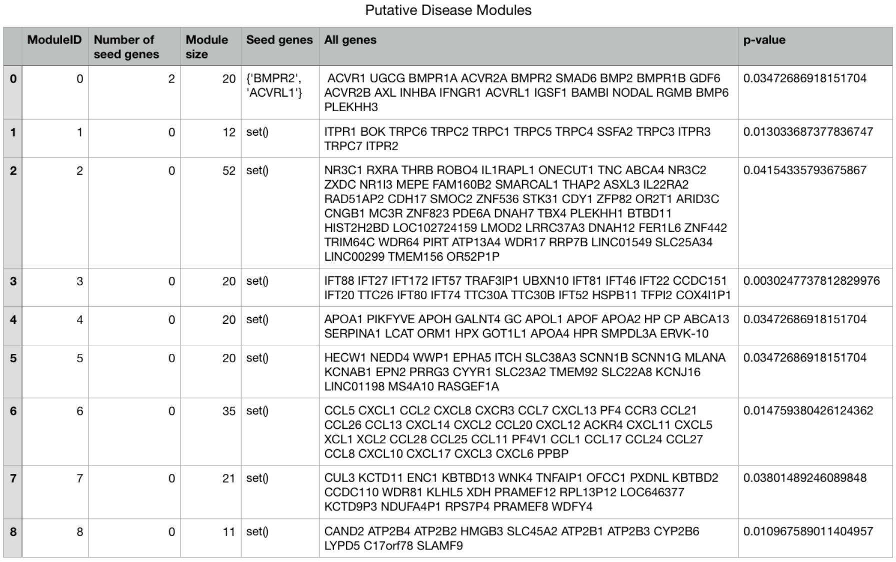
t	Recall@50	Recall@25	Recall@15	Recall@1
0.01	0.1586	0.0897	0.0516	0.0034
0.002	0.1552	0.0897	0.0516	0
0.005	0.1552	0.08979	0.0516	0.0034

t	F1_score@50	F1_score@25	F1_score@15	F1_score@1
0.01	0.1706	0.1256	0.0822	0.0066
0.002	0.1668	0.1256	0.0822	0
0.005	0.1668	0.1256	0.0822	0.0066

Figure 4: Performances for Diffusion Based Algorithm for different t

As we can see from the tables the diffusion based algorithms with $t = 0.01$ outperforms the other methods.

For what concern the putative gene identification we obtained the following results:



Putative Disease Modules

	ModuleID	Number of seed genes	Module size	Seed genes	All genes	p-value
0	0	2	20	{'BMPR2', 'ACVRL1'}	ACVR1 UGCG BMPR1A ACVR2A BMPR2 SMAD6 BMP2 BMPR1B GDF6 ACVR2B AXL INHBA IFNGR1 ACVRL1 IGSF1 BAMBI NODAL RGMB BMP6 PLEKHH3	0.03472686918151704
1	1	0	12	set()	ITPR1 BOK TRPC6 TRPC2 TRPC1 TRPC5 TRPC4 SSFA2 TRPC3 ITPR3 TRPC7 ITPR2	0.013033687377836747
2	2	0	52	set()	NR3C1 RXRA THRB ROBO4 IL1RAPL1 ONECUT1 TNC ABCA4 NR3C2 ZXDC NR1I3 MEPE FAM160B2 SMARCAL1 THAP2 ASXL3 IL22RA2 RAD51AP2 CDH17 SMOC2 ZNF536 STK31 CDY1 ZFP82 OR2T1 ARID3C CNGB1 MC3R ZNF823 PDE6A DNAH7 TBX4 PLEKHH1 BTBD11 HIST2H2BD LOC102724159 LMOD2 LRRC37A3 DNAH12 FER1L6 ZNF442 TRIM64C WDR64 PIRT ATP13A4 WDR17 RRP7B LINC01549 SLC25A34 LINC00299 TMEM156 OR52P1P	0.04154335793675867
3	3	0	20	set()	IFT88 IFT27 IFT172 IFT57 TRAF3IP1 UBXN10 IFT81 IFT46 IFT22 CCDC151 IFT20 TTC26 IFT80 IFT74 TTC30A TTC30B IFT52 HSPB11 TFPI2 COX4I1P1	0.0030247737812829976
4	4	0	20	set()	APOA1 PIKFYVE APOH GALNT4 GC APOL1 APOF APOA2 HP CP ABCA13 SERPINA1 LCAT ORM1 HPX GOT1L1 APOA4 HPR SMPDL3A ERVK-10	0.03472686918151704
5	5	0	20	set()	HECW1 NEDD4 WWP1 EPHA5 ITCH SLC38A3 SCNN1B SCNN1G MLANA KCNAB1 EPN2 PRRG3 CYYR1 SLC23A2 TMEM92 SLC22A8 KCNJ16 LINC01198 MS4A10 RASGEF1A	0.03472686918151704
6	6	0	35	set()	CCL5 CXCL1 CCL2 CXCL8 CXCR3 CCL7 CXCL13 PF4 CCR3 CCL21 CCL26 CCL13 CXCL14 CXCL2 CCL20 CXCL12 ACKR4 CXCL11 CXCL5 XCL1 XCL2 CCL28 CCL25 CCL11 PF4V1 CCL1 CCL17 CCL24 CCL27 CCL8 CXCL10 CXCL17 CXCL3 CXCL6 PPBP	0.014759380426124362
7	7	0	21	set()	CUL3 KCTD11 ENC1 KBTBD13 WNK4 TNFAIP1 OFCC1 PXDNL KBTBD2 CCDC110 WDR81 KLHL5 XDH PRAMEF12 RPL13P12 LOC646377 KCTD9P3 NDUFA4P1 RPS7P4 PRAMEF8 WDFY4	0.03801489246089848
8	8	0	11	set()	CAND2 ATP2B4 ATP2B2 HMGB3 SLC45A2 ATP2B1 ATP2B3 CYP2B6 LYPD5 C17orf78 SLAMF9	0.010967589011404957

Figure 5: Putative Disease Genes

With network procedures we obtained potential genes that can be linked with the disease. Most interesting case is the first one where we found two genes that we already know that are linked with the disease.

Now let's take a look to the results of our enrichment analysis, specifically looking at the plots created using EnrichR and commenting them:

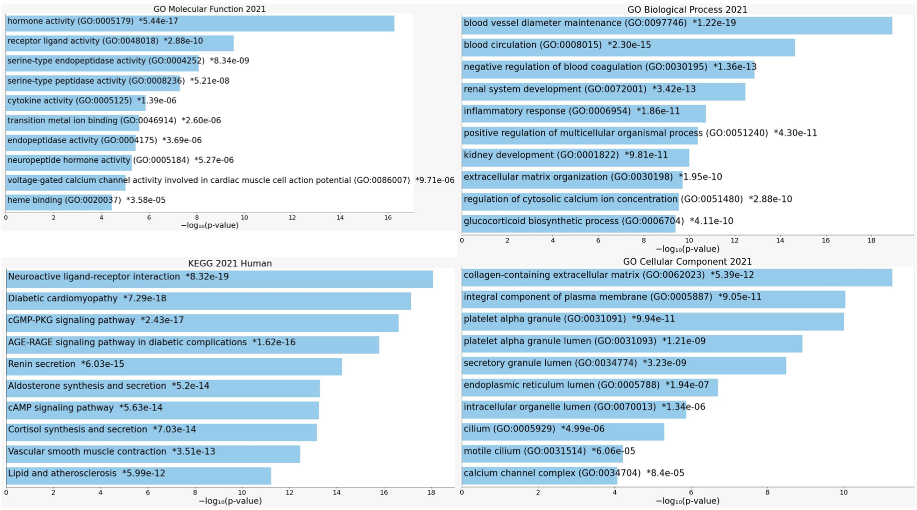


Figure 6: Enrichment Analysis

Starting from the GO-MF plot we can see the highest relevance associated to hormone activity. This result can also been found in the literature, see: [1].

In the next one GO-BP we can see the highest relevance with blood vessel diameter maintenance. This result is clearly related with Hypertension considering that the diameter of vessels affects blood pressure.

In the bottom left, KEGG, we can see the highest relevance associated with Neuroactive Ligand-Receptor Interaction. This result is confirmed and justified by literature, e.g. [2]

In bottom right we have GO-CC and the highest relevance is associated with Collagen Containing Extracellular Matrix. This result is again justified by literature, e.g. [3].

For what concern the last part we used ”clinicaltrials.gov” [12] in order to find the number of studies related to each of the drugs that we saw occurring the most with DGIdb dataset we downloaded from the website; we found that the drugs occurring the most are TAMOXIFEN, IMIPRAMINE, AMOXAPINE, DESIPRAMINE, PROTRIPTYLINE, PSEUDOEPHEDRINE all occurring with the same frequency. For the first one we found 3 clinical trials in relation with Hypertensive Disease. We found 23 studies for PSEUDOEPHEDRINE. There aren’t studies for the other drugs in relation with the disease.

After all this results we can make the following conclusions:

- The best algorithm, between the ones that we used and in terms of performances, is the Diffusion Based algorithm, specifically the one with a diffusion time $t = 0.01$.

- Considering the first 200 putative genes obtained with the best algorithm (Diffusion Based) we performed the Enrichment Analysis. The results we obtained are positive in terms of studies, in the sense that the conditions that are significative in relation with our genes for our analysis already appear in literature

- From what concern the drugs associated to the disease that we found in the last part seems that two of them are already in literature w.r.t. to our disease, while other don't. We don't have the instrument to assess anything more specific but we can suppose that, as the ones that have published studies, the ones that don't have studies associated are possible suggestion for drugs to affect (optimistically affect positively) the disease that we studied.

Author Contributions

S. Boesso in [8] he carried out task 4 and contributed to task 2, L. Lavagna in [8] he carried out task 1, 2, and 3 and developed the auxiliary functions, L. Skerl in [8] carried out task 5 and he mainly wrote the report.

References

[1] Ladan Mehran, Negar Delbari, Atieh Amouzegar, Mitra Hasheminia, Maryam Tohidi, Fereidoun Azizi, Reduced Sensitivity to Thyroid Hormone Is Associated with Diabetes and Hypertension, The Journal of Clinical Endocrinology and Metabolism, Volume 107, Issue 1, January 2022, Pages 167–176

[2] Zhang Q, Yang J, Yang C, Yang X, Chen Y. Eucommia ulmoides Oliver-Tribulus terrestris L. Drug Pair Regulates Ferroptosis by Mediating the Neurovascular-Related Ligand-Receptor Interaction Pathway- A Potential Drug Pair for Treatment Hypertension and Prevention Ischemic Stroke. Front Neurol. 2022 Mar 8;13:833922. doi: 10.3389/fneur.2022.833922. PMID: 35345408; PMCID: PMC8957098.

[3] Ponticos M, Smith BD. Extracellular matrix synthesis in vascular disease: hypertension, and atherosclerosis. J Biomed Res. 2014 Jan;28(1):25-39. doi: 10.7555/JBR.27.20130064. Epub 2013 Sep 20. PMID: 24474961; PMCID: PMC3904172.

[4] Theresa A. Gelzinis, Pulmonary Hypertension in 2021: Part I—Definition, Classification, Pathophysiology, and Presentation, Journal of Cardiothoracic and Vascular Anesthesia, Volume 36, Issue 6, 2022, Pages 1552-1564, ISSN 1053 0770

[5] `https://github.com/dinaghiassian/DIAMOnD`

[6] `https://micans.org/mcl/`

[7] `https://maayanlab.cloud/Enrichr/`

[8] `https://github.com/leonardoLavagna/Drug-repurposing`

[9] `https://cytoscape.org`

[10] `https://thebiogrid.org`

[11] www.dgidb.org

[12] https://clinicaltrials.gov