Putative disease gene identification and drug repurposing for retinal dystrophy

Bioinformatics and Network Medicine - M.Sc. Data Science - Sapienza Università di Roma A.Y. 2024-25

F.Mari 1919565, L.Oddi 1846085, L.Pannacci 1948926
Group 6

January 19, 2025

Abstract

In this study we analyze a human protein-protein interaction database and a gene-disease association database for retinal dystrohpy, perform a comparative analysis with cross-validation to find the best performing algorithm between three state-of-the-art disease-gene identification algorithms (DIAMOnD, DiaBLE, diffusion-based) and apply it to find putative disease genes. An enrichment analysis is performed to find overrepresented functions and the putative disease genes are used to identify already-approved drugs potentially effective against the disease. Finally we study wether those drugs have already been proposed to combat the disease of interest and are being used in clinical trials.

While the evidence gathered is weak probably due to the structure of the GDA, which itself is most likely caused by the well-known genetic heterogeneity of the disease, Duloxetine Hydrochloride and Betaine Hydrochloride are found as drug candidates and their influenced pathways seems to be indeed related to retinal conditions, but no clinical trials involving them and retinal dystrohpy currently exist.

Introduction

Putative disease gene identification via network-based methods is one of the main challenges in network medicine. The underlying hypothesis is that members of the same disease module share topological and functional features. The availability of PPI networks (a.k.a. interactomes), among other data, allows the use graph-based algorithms to infer new putative genes-disease associations (GDAs). Several tools have been developed with this aim following different approaches: Markov Clustering [1] uses as main driver the hypothesis that topological modules based on link density are the best predictor for disease modules. Other algorithms such as DIAMOnD [2] and DiaBLE [3] instead focus on connectivity significance. Another class of algorithms is diffusion-based, either using random walks with restart [4] [5], heat diffusion [6] or more recently quantum walks [7].

Another important field of network medicine is network-based drug repurposing, i.e. computational approaches that aim to find new uses for already-approved drugs, using biological data such as drug-disease association graphs [8]. While classic approaches for drug repurposing are experimental and often based on the reported side-effects of drug molecules, computational drug repurposing exploits knowledge of disease genes, biological pathways and drug targets to propose new clinical trials, effectively skipping the first phases of the drug development process, reducing noticeably both time and costs [9]. For those reasons drug repurposing is strongly supported by governments and large public databases are available [8].

Retinal dystrophies are degenerative and hereditary eye conditions that affect the retina and the choroid, causing a wide variety of symptoms including contraction of the visual field, night blindness, light sensitivity, blurred vision, alteration in color perception and usually progress towards complete blindness [10]. Medical literature agrees on the high clinical and genetic heterogeneity of the disease [10] [11] [12] [13], that is, many genes seem to be related to the disease and different mutations of the same gene seem to cause different disease phenotypes. Literature also agrees that currently no cure exists for the disease, with gene therapy seeming the most promising of possible treatments.

Our study include both the fields described above with the aim of firstly find new putative disease genes for retinal dystrophy using some of the aforementioned algorithms, infer from them pathways and molecular functions via enrichment analysis and finally propose the use of new drugs on the basis of the newfound knowledge.

Materials and Methods

In this section we explain and discuss the experimental procedures and methods used. Following the homework's directives the study has been divided into four subsections plus a fifth optional task. All coding parts of the study have been written in Python and are available as an attachment to the same mail the report was attached to.

Data gathering and interactome reconstruction

For our analysis we exploit two datasets, a PPI network provided by BioGRID [14] (version 4.4.240) and a GDA database for retinal dystrophy gathered via DisGeNET [15] (provided on the Classroom page of the course).

We process the PPI data by selecting only human genes, i.e. only rows where both 'organism A' and 'organism B' fields have identifier '9606' for Homo Sapiens. Then we filter out non-physical interactions and remove self-loops. Lastly we isolate the largest connected component of the network. On this network we perform a topological analysis to get some intel about its structure.

For the GDA data we perform a cleanup of all non 'protein-coding' genes (this passage is not explicitly stated in the homework text, but it has been suggested by professor Tieri during a lecture) and check the symbols using HGNC's multi-symbol checker tool [16] (available at the following link) and to solve conflicting information a manual check is performed using the other fields available in the database. The final step is to isolate the largest connected component of the disease interactome, i.e. the subgraph of the PPI LCC where only disease genes are considered. Finally for the genes present in the GDA data but not in the PPI LCC we perform a failure analysis to determine the reasons for their absence and exclude them from the analysis.

Comparative analysis of disease gene identification algorithms

To find new putative disease genes three different algorithms have been tested out. All of them require the same data: a network of protein-protein interactions and a set of already known disease genes. At the end of the study a fourth

algorithm has been tested out as an optional task. A comparative analysis is performed to choose the best out of the first three algorithms by performing a **5-fold cross validation** to test the ability of the algorithms to retrieve one fifth of the seed set by using as input the rest.

DIAMOnD [2] (DIseAse MOdule Detection) is a popular module-based algorithm for the proposal of putative disease genes (a.k.a. gene prioritization tool) whose innovation is the use of *connectivity significance* instead of link density to characterize disease modules. The algorithm works by iteratively computing *hypergeometric tests* for all candidate genes against the null hypothesis that disease genes are randomly scattered across the interactome and the candidate with the lowest p-value is added to the seed set. The p-values are computed as follows:

$$\text{p-value} = \sum_{k_i=k_s}^{k} p(k, k_i) \tag{1}$$

With:

$$p(k, k_s) = \frac{\binom{s_0}{k_s}\binom{N-s_0}{k-k_s}}{\binom{N}{k}} \tag{2}$$

Where Equation 2 is the PMF of the hypergeometric distribution, that is, the probability for a gene with k links to have k_s of them with seed genes, given a PPI network of size N and a seed set of size s_0, under the null hypothesis. The code for the algorithm is downloaded from the original repository and only minor changes has been performed to avoid deprecation warnings.

DiaBLE [3] (DIAMOnD Background Local Expansion) is a modification of DIAMOnD that introduces the concept of *dynamic gene universe*: instead of using the whole interactome as gene universe, DiaBLE uses the set composed of seed genes, candidate genes (i.e. those having at least one link to the seed set) and the first neighbors of the candidates. The only major difference in the implementation is that the value of N in Equation 2 changes every time a new gene is added and must be updated for each iteration. The implementation is a modification of the DIAMOnD code where the changing universe size has been implemented.

The third is a **diffusion-based algorithm**, it uses *Fourier's laws of heat conduction* as a way to test the distance between a node and a sets of nodes. This can be seen equivalently as a random walk on an undirected graph without restarts and with time steps approaching zero. The formula governing heat diffusion can be obtained solving the following differential equation:

$$\frac{dh(t)}{dt} = -Lh(t) \tag{3}$$

Which is solved by:

$$h(t) = e^{-Lt}h(0) \tag{4}$$

Where L is the graph Laplacian and $h(0)$ the initial heat configuration. The implementation of the algorithm is adapted from a laboratory made during the course that was based on the Cytoscape Heat Diffusion service, which repository is available here. The only significative change made to the algorithm is to make it return also the heat value associated with each node.

PROCONSUL [17] (PRObabilistic exploration of CONnectivity Significance patterns for disease modULe discovery) is another modification of DIAMOnD which exploits the idea of *probabilistic exploration* of the space of putative disease genes with the objective of finding paths that may seem less relevant at first but lead to overall better results. It modifies DIAMOnD by converting the output of the hypothesis tests into a probability distribution by applying a soft-max over the negative logarithms of the p-values and drawing the new seed gene from this distribution instead of selecting always the element with the lowest p-value. To reduce statistic fluctuations and enhance accuracy techniques like temperature, the averaging of results from multiple runs and different sampling strategies (e.g. nucleus sampling) are applied. The algorithm has been downloaded from the original repository and adapted to the notebook.

In order to compare the algorithms we chose three different metrics: precision, recall and F1-score. When comparing algorithms for putative disease gene identification, selecting the best algorithm should ideally consider the fact that false negatives can have significant consequences in disease research. Since **recall** is important for capturing as many true disease-related genes as possible, it might be the most important metric. **Precision** is also relevant because falsely identifying irrelevant genes as disease-related could waste resources and lead to incorrect conclusions. **F1-score** is a good metric if we want to balance precision and recall.

Putative disease gene identification

We choose the best performing algorithm according to the previous metrics and predict a list of 100 putative disease genes, this time using all known GDAs as seed genes and an enrichment analysis is then performed on both the original seed set and the newfound putative disease genes.

An **enrichment analysis** is a statistical method used to identify biological processes, pathways, or functions that are overrepresented (i.e. enriched) in a given set of genes, compared to a background set. It helps in interpreting large gene lists by associating them with known biological knowledge, such as Gene Ontology (GO) terms (e.g., biological processes, cellular components, molecular functions) or pathways (e.g., KEGG's, Reactome's). Similarly to DIAMOnD this is also usually implemented with hypergeometric tests.

To perform the enrichment analysis both genes sets are inserted in the EnrichR software (available here) and tables regarding GO-Biological Processes (GO-BP), GO-Molecular Functions (GO-MF), GO-Cellular Components (GO-CC), Reactome Pathways, and KEGG Pathways are downloaded. Finally the overlap between the enriched functions of the two sets is evalued, after selecting only the significative terms (i.e. associated with *adjusted p-value* < 0.05).

Drug repurposing

To identify potential drug candidates for repurposing, a list of putative disease genes is curated from the rankings obtained in the previous analysis; specifically, the top 20 genes from the ranking generated at point 3.1 are selected for further analysis. A database of drug-gene interaction data is retrieved from DGIdb (available here); it is used the latest available version of the *interactions.tsv* file and the genes of interest are cross-referenced with this database to identify approved drugs targeting these genes.

A filtering of the database is performed by selecting rows where the gene name matched the putative genes and only drugs approved for human use are kept by filtering entries where the 'approved' column was set to 'True'. The identified drugs are then ranked according to the number of putative disease genes they are associated with to emphasize drugs with greater relevance to the prioritized gene set and

the top three drugs from the ranking are validated by querying ClinicalTrials.gov for their involvement in clinical trials targeting the disease of interest. The search is conducted using 'Retinal dystrophy' in the 'Condition or Disease' field, 'Clinical trial' in the 'Other terms' field and the proposed drug in the 'Intervention or treatment' field (Figure 7).

PROCONSUL comparison

The PROCONSUL algorithm is executed with a temperature set to temp=1, and the top 20 genes from the resulting ranked list are selected for comparison with the top 20 genes identified by the best-performing algorithm used in point 3.1. To evaluate the similarity between these ranked lists, the overlap between the two sets of genes was assessed by performing an intersection of the gene names. Additionally, we propose two other metrics for the comparison:

• **Weighted Kendall's Tau** [18]: metric which evaluates the agreement (i.e. correlation) between rankings; this weighted version exploits both the relative rankings and the scores obtained from the algorithms to introduce information about the difference in scores between elements. Defined as:

$$\tau_w = \frac{\sum_{i<j} w_{ij} \cdot \text{concordant}(i,j)}{\sum_{i<j} w_{ij}}$$

Where $\text{concordant}(i,j) = 1$ iff the relative ranking of the two elements is the same and 0 otherwise and w_{ij} a weight function that we set to be reciprocate of the absolute difference between the normalized scores of the two algorithms.

• **Spearman's Footnote Distance** [19]: distance metric which measures the sum of absolute differences of ranks between corresponding items, offering a straightforward numerical measure of rank deviation. Defined as:

$$H(R_1, R_2) = \sum_{i=1}^{n} |R_1(i) - R_2(i)|$$

With R_1 and R_2 the two rankings, n the amount of elements to be ranked and $R_j(i)$ the position of item i in ranking j.

Results and Discussion

In this section we describe the obtained results, discuss and interpret them. As for 'Material and Methods' this section is also divided into four subsections plus one for the optional task. All the figures and tables that will be referenced are found at the end of the report.

Data gathering and interactome reconstruction

The PPI database records 1.265.586 interactions of which 94.104 include non-human proteins and are removed, an ulterior 18.698 entries are removed due to being non-physical and 7.976 self-loops are also discarded. We end up with 1.144.808 interactions and instead of manually filtering repeated edges we exploit the structure of the NetworkX *Graph* class to ignore them and end up with a graph having 19.972 nodes (genes) and 861.240 edges (physical interactions). Interestingly enough the network is already a single connected component.

To analyze the topology we perform a **power-law fit** to check if the network follows a scale-free structure. The fit is performed with a MLE method and goodness-of-fit is measured with a Kolmogorov-Smirnov test [20], which returns a KS statistic of 0.02243 with a p-value of 0.8054, suggesting the network indeed follows a power-law distribution. A plot of the power-law fit is included in the report (Figure 1). A low clustering coefficient of 0.1363 make us exclude a small-world topology. The plot of the subgraph of nodes with degree $\geq$ 2000 (Figure 2) shows how well interconnected the network is. All those observations will be very important in the analysis of the results we got in the comparative analysis of the gene identification algorithms.

The GDA database records 263 genes of which 47 are discarded due to being not protein-coding and for the rest symbol checking is performed. While all inputted symbols are 'Approved' symbols, some are also 'Alias' or 'Previous' symbols of other genes and for them the manual check has been performed. The construction of the disease interactome gave us a network with 199 nodes and 140 edges.

For the 17 missing genes we have performed a failure analysis to determine wether our symbol checking has been incorrect and noticed that 2 of them are present in the interactome but their interactions are not physical and have been discarded while the remaining are all genes for which HGNC only returned an 'Approved' symbol and were therefore left unchanged during the symbol check.

The disease interactome is composed of a single large connected component, few little components with 2-3 nodes and a majority of unconnected nodes (Figure 3). Selecting the LCC we end up with a graph of 72 nodes and 120 edges (Figure 4). We obtained a small LCC, about 1/3 of the size of the GDA. Confronting this value to what is obtained using the datasets assigned to other teams a large difference is evident, with their LCC being always a larger proportion of the dataset, usually about 150-200 genes and also much better connected. We believe having a disease interactome so scattered can have a very negative impact in the performances of the disease-gene identification algorithms.

The requested metrics are available in Table 1 and 2 and the node degree vs. node betweenness scatterplot in Figure 5. As consequence of the structure of our GDA both the metrics recorded in the tables and the scatterplot are different from what could be expected.

Comparative analysis of disease gene identification algorithms

Algorithms have been modified as described in 'Materials and Methods' and a random seed has been set for the cross-validation split to ensure the repeatability of the results. The requested metrics are reported for all algorithms using different amounts of selected genes: 50, $\frac{1}{10}n$, $\frac{1}{4}n$, $\frac{1}{2}n$, n with n the number of disease seed genes and for the diffusion algorithm three diffusion times have been used: 0.002, 0.005, 0.01. Results for DIAMOnD are reported in Table 3, for DiaBLE in Table 4 and for diffusion in Table 5.

Confronting the results of DIAMOnD and DiaBLE we can notice that DiaBLE gets better results in every metric and for any amount of genes recovered. However if we change the input seed set from just the LCC to all the disease genes the results of the two algorithms become exactly the same. Studying the motivations for this behavior we found that using all the disease genes as seed set the amount of candidates is 6.239 and adding also their neighbors we get a DiaBLE universe of 19.700 for the first iteration, which is 98.64% of the whole interactome. If the same is done on the seed set used in the cross-validation the amount of candidates drops to 4.075 and the total universe size to 19.359. While this change seems small it is enough to significantly change the results observed in the metrics. We

believe that the reasons for this large universe size must be attributed to the topology of the network, which was studied above. The scale-free topology implies that if even only one of our disease genes is a hub or is connected to a hub the size of DiaBLE's universe is bound to explode.

Another thing worth noting is that **the three proposed diffusion times for the diffusion algorithm lead to no change in the ranking**. We can also attribute this behavior to the network topology: being so interconnected the relative rankings of genes quickly reach a somewhat stable state. While it is true that the ranking remains unchanged, observing the heat values of each node we find that they continue to increase with time as the seed genes nodes still have a higher heat value. While this could lead to think that increasing the diffusion time could help get better results we believe that this would instead be detrimental, as we would lose the locality significance of the results and end up in a globally stable state that doesn't take into account the starting heat configuration. In general on a single connected component for $t \to \infty$ each node reaches a heat value of $\frac{s_0}{N}$, with s_0 size of the seed set and N size of the component.

Putative disease gene identification

The **diffusion-based algorithm** was found to be the best algorithm for the amount of genes we want with respect to all the three metrics discussed in the 'Methods' section. The 100 putative disease genes were then predicted with that algorithm using a diffusion time of $t = 0.002$. A plot of the putative disease module is reported in Figure 6. After that the biological validation was performed through the enrichment analysis between the original list of seed genes and the list of 100 genes predicted by the diffusion algorithm: EnrichR results are available for the original genes and the putative genes. Only terms associated with adjusted p-value < 0.05 were retrieved for further analysis.

The highest amount of overlapping terms was found in the Reactome Pathways with 7 terms, corresponding to the 2% of the total. We also found 3 overlapping terms with on Cellular Components (1% in proportion), while nothing was found with respect to the KEGG Pathways, Molecular Function or Cellular Component.

Drug repurposing

From the 20 putative disease genes identified, only two genes (**BHMT** and **NRXN1**) were linked to approved drugs (**Betaine Hydrochloride**, **Duloxetine Hydrochloride**, **Betaine**, and **Nicotine Polacrilex**) present in the DGIdb database. However, none of these drugs are directly related to retinal dystrophy nor are they currently being tested in clinical trials for this condition.

The absence of BHMT and NRXN1 in the gene-disease association database further suggests that these genes are not currently known to be directly associated with retinal dystrophy. Since a GDA database typically contains curated gene-disease associations derived from biomedical literature, coming from either experimental studies or computational proofs, their absence in this dataset could indicate that:

1) These genes have not been studied or reported in connection with retinal dystrophy.
2) There might be emerging researches implicating the genes in relevant pathways, but they haven't reached a level of confidence for their inclusion in the database.
3) The genes might influence pathways or processes relevant to retinal dystrophy, but they are not directly implicated.

Although the absence of these genes in the GDA dataset suggests a weak or indirect link to retinal dystrophy, it does not entirely rule out their relevance. Further exploration, such as pathway analysis, is necessary to investigate if these genes are part of broader mechanisms influencing the disease. Moreover, the relevance of drugs targeting these genes (e.g., Betaine and Duloxetine) may be more exploratory, given the weak gene-disease connections.

The limited overlap between predicted genes and known drug targets reveals significant challenges in the study of retinal dystrophy and could be linked to the high genetic heterogeneity that characterize the disease discussed in medical literature, hindering the use of computational approaches that are purely network-based and don't take into account ulterior features of the genes. Those characteristics of the disease are reflected in the disease interactome studied during 'Data gathering', where we found that the largest connected component had few genes while the majority was in small clusters or completely unconnected.

While the found drugs are not currently considered candidates for treating retinal dystrophy, their pharmacological properties may offer indirect benefits:

• **Duloxetine Hydrochloride**, a serotonin-norepinephrine reuptake inhibitor (SNRI), is primarily prescribed for depression and neuropathic pain. Its mechanism of action includes modulating central nervous system activity and influencing inflammatory pathways [21]. Neuroinflammation has been implicated as a contributing factor in retinal dystrophies, specifically in the degeneration of photoreceptors and retinal ganglion cells [22]. While further research is needed, Duloxetine's effects on inflammatory pathways may provide insights into its potential relevance for retinal conditions.

• **Betaine and derivatives** (*Betaine Hydrochloride in this case*), are involved in methylation and homocysteine metabolism, which are crucial for maintaining cellular homeostasis and repair mechanisms. Disruptions in these pathways have been linked to oxidative stress and cell death, processes also implicated in retinal degeneration [23] [24].

While the identified drugs are not immediate candidates for retinal dystrophy treatment, their pharmacological properties could provide insights into alternative mechanisms or therapies that may be worth exploring in future studies.

PROCONSUL comparison

The **PROCONSUL algorithm** produced a ranked list of genes based on their association with the disease module, but the comparison between the lists produced by the two algorithms was limited due to a minimal overlap, with only one common gene identified (**LZTFL1**). Despite the lists being of the same size, the lack of substantial overlap made it challenging to draw meaningful conclusions using the two intended comparison metrics.

In addition to this analysis, the ranked list of genes produced by PROCONSUL was further explored for drug repurposing opportunities. Unfortunately, this effort was inconclusive, as no approved drugs targeting the PROCONSUL-identified genes were found in the DGIdb database. We believe that those two results are also linked to the 'difficulty' of the assigned disease, which was discussed in the previous section.

Nevertheless, these methods might still be useful in identifying disease-related genes when there is greater concordance between the rankings.

Authors Contribution

F.Mari:
Code: GDA gathering, symbol checking, collection of network metrics of disease LCC genes and GDAs basic network data, Enrichment Analysis (3.2).
Report: Introduction, Materials and Methods (Data gathering, Putative disease gene identification), Results (Putative disease gene identification). Figure 5.

L.Oddi:
Code: Putative disease gene identification (3.1), Drug repurposing, optional task (PROCONSUL). Descriptive and theory part comments in the code.
Report: Materials and Methods (Drug repurposing, PROCONSUL), Results (Drug repurposing, PROCONSUL). Table 1 and 2. Figure 7. References (.bib file).

L.Pannacci:
Code: PPI network cleaning, construction and topology analysis, missing genes analysis, modification and implementation of disease-gene identification algorithms and validation. Plotting and analysis comments in the code.
Report: Abstract, Introduction, Material and Methods (Comparative analysis), Results (Data gathering, Comparative analysis), Report pagination. Tables 3, 4, 5; Figures 1, 2, 3, 4, 6.

References

[1] Stijn Van Dongen. "Graph Clustering Via a Discrete Uncoupling Process". In: *SIAM Journal on Matrix Analysis and Applications* 30.1 (2008), pp. 121–141. DOI: 10.1137/040608635.

[2] Ghiassian et al. "A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome". In: *PLOS Computational Biology* 11.4 (Apr. 2015), pp. 1–21. DOI: 10.1371/journal.pcbi.1004120.

[3] Manuela Petti et al. "Connectivity significance for disease gene prioritization in an expanding universe". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 17.6 (2019), pp. 2155–2161. DOI: 10.1109/TCBB.2019.2938512.

[4] Sebastian Kohler et al. "Walking the interactome for prioritization of candidate disease genes". In: *The American Journal of Human Genetics* 82.4 (2008), pp. 949–958. DOI: 10.1016/j.ajhg.2008.02.013.

[5] Alberto Valdeolivas et al. "Random walk with restart on multiplex and heterogeneous biological networks". In: *Bioinformatics* 35.3 (Aug. 2018), pp. 497–505. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty637.

[6] Daniel E. Carlin et al. "Network propagation in the cytoscape cyberinfrastructure". In: *PLOS Computational Biology* 13.10 (Oct. 2017), pp. 1–9. DOI: 10.1371/journal.pcbi.1005598.

[7] Harto Saarinen et al. *Disease Gene Prioritization With Quantum Walks*. 2023. arXiv: 2311.05486 [quant-ph]. URL: https://arxiv.org/abs/2311.05486.

[8] T. N. Jarada et al. "A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions". In: *Journal of Cheminformatics* 12.1 (2020), pp. 46–46. DOI: 10.1186/s13321-020-00450-7.

[9] V. S. Kulkarni et al. "Drug Repurposing: An Effective Tool in Modern Drug Discovery". In: *Russian Journal of Bioorganic Chemistry* 49.2 (2023). Epub 2023 Feb 21, PMID: 36852389, PMCID: PMC9945820, pp. 157–166. DOI: 10.1134/S1068162023020139.

[10] B. M. Nash et al. "Retinal dystrophies, genomic applications in diagnosis and prospects for therapy". In: *Translational Pediatrics* 4.2 (Apr. 2015). PMID: 26835369, PMCID: PMC4729094, pp. 139–163.

[11] Murro et al. "Multidisciplinary approach to inherited retinal dystrophies from diagnosis to initial care: review with inputs from clinical practice". In: *Orphanet Journal of Rare Diseases* 18 (2023), p. 223. DOI: 10.1186/s13023-023-02798-z.

[12] M. Talib and C. J. F. Boon. "Retinal Dystrophies and the Road to Treatment: Clinical Requirements and Considerations". In: *Asia-Pacific Journal of Ophthalmology (Phila)* 9.3 (May 2020). PMID: 32511120, PMCID: PMC7299224, pp. 159–179. DOI: 10.1097/APO.0000000000000290.

[13] Marina França Dias et al. "Molecular genetics and emerging therapies for retinitis pigmentosa: Basic research and clinical perspectives". In: *Progress in Retinal and Eye Research* 63 (2018), pp. 107–131. ISSN: 1350-9462. DOI: 10.1016/j.preteyeres.2017.10.004.

[14] R. Oughtred et al. "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions". In: *Protein Science* 30.1 (2021), pp. 187–200. DOI: 10.1002/pro.3978.

[15] J. Piñero et al. "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants". In: *Nucleic Acids Research* 45.D1 (2017), pp. D833–D839. DOI: 10.1093/nar/gkw943. eprint: 2016Oct19.

[16] R.L. Seal et al. "Genenames.org: the HGNC resources in 2023". In: *Nucleic Acids Research* (2023). Accessed on January 2025. DOI: 10.1093/nar/gkac888.

[17] Riccardo De Luca et al. "PROCONSUL: PRObabilistic exploration of CONnectivity Significance patterns for disease modULe discovery". In: (2022), pp. 1941–1947. DOI: 10.1109/BIBM55620.2022.9995586.

[18] Grace S. Shieh. "A weighted Kendall's tau statistic". In: *Statistics and Probability Letters* 39.1 (1998), pp. 17–24. ISSN: 0167-7152. DOI: 10.1016/S0167-7152(98)00006-6.

[19] Diaconis et al. "Spearmans Footrule as a Measure of Disarray". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.2 (1977), pp. 262–268.

[20] M. L. Goldstein, S. A. Morris, and G. G. Yen. "Problems with fitting to the power-law distribution". In: *The European Physical Journal B* 41.2 (Sept. 2004), pp. 255–258. ISSN: 1434-6036. DOI: 10.1140/epjb/e2004-00316-5.

[21] DrugBank. *Duloxetine (DB00476)*. https://go.drugbank.com/drugs/DB00476. 2025.

[22] Maria H. et al. "Contribution of Microglia-Mediated Neuroinflammation to Retinal Degenerative Diseases". In: *Mediators of Inflammation* 2015 (2015). Accessed: 2025-01-09, p. 673090. DOI: 10.1155/2015/673090.

[23] DrugBank. *Betaine (DB06756)*. https://go.drugbank.com/drugs/DB06756. 2025.

[24] DrugBank. *Betaine Hydrochloride (DBSALT001448)*. https://go.drugbank.com/salts/DBSALT001448. 2025.

Tables

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
Retinal Dystrophy	C0854723	C11	263	199	72

Table 1: Summary of GDAs and basic network data.

Ranking	Gene name	Degree	Betweenness	Eigenvector centrality	Closeness centrality	Ratio Betw./Degree
1	OFD1	11	0.0342	0.3758	0.2367	0.0031
2	BBS7	11	0.0120	0.1322	0.1888	0.0011
3	PRPF8	9	0.0238	0.2146	0.2351	0.0026
4	BBS4	8	0.0220	0.1760	0.2052	0.0028
5	BBS1	8	0.0036	0.0285	0.1766	0.0004
6	PRPF4	7	0.0061	0.0427	0.2139	0.0009
7	PRPF6	7	0.0133	0.0933	0.2305	0.0019
8	PRPF3	7	0.0061	0.0427	0.2139	0.0009
9	PRPF31	7	0.0356	0.2495	0.2399	0.0051
10	SNRNP200	7	0.0053	0.0369	0.2152	0.0008
11	BBS2	7	0.0	0.0	0.1762	0.0
12	BBS5	6	0.0	0.0	0.1757	0.0
13	BBS9	6	0.0	0.0	0.1757	0.0
14	TTC8	6	0.0	0.0	0.1757	0.0
15	RPGRIP1L	6	0.0183	0.1101	0.2268	0.0031
16	CEP250	5	0.0141	0.0703	0.2101	0.0028
17	RPGR	5	0.0512	0.2561	0.2276	0.0102
18	ZNF408	4	0.0899	0.3596	0.2491	0.0225
19	PAX2	4	0.0	0.0	0.1749	0.0
20	CEP164	4	0.0268	0.1071	0.2152	0.0067
21	LCA5	4	0.0207	0.0829	0.2082	0.0052
22	NPHP4	4	0.0187	0.0746	0.2145	0.0047
23	DHX38	4	0.0140	0.0559	0.1935	0.0035
24	GIGYF2	4	0.0644	0.2575	0.2367	0.0161
25	IQCB1	4	0.0128	0.0513	0.2023	0.0032
26	IDH3B	3	0.0004	0.0013	0.2082	0.0001
27	IFT140	3	0.0016	0.0049	0.1715	0.0005
28	TTC21B	3	0.0157	0.0472	0.1972	0.0052
29	EMC1	3	0.0443	0.1328	0.1967	0.0148
30	RP2	3	0.0276	0.0829	0.1458	0.0092
31	LRP2	3	0.0685	0.2056	0.2094	0.0228
32	TIMP3	3	0.0530	0.1590	0.1788	0.0177
33	ITGA4	3	0.0555	0.0555	0.1935	0.0185
34	RPGRIP1	3	0.0293	0.0293	0.2219	0.0098
35	CEP290	3	0.0048	0.0145	0.2058	0.0016
36	AHI1	3	0.0048	0.0145	0.2058	0.0016
37	CC2D2A	3	0.0000	0.0001	0.1940	0.0000
38	BBS12	2	0.0141	0.0282	0.1599	0.0070

Ranking	Gene name	Degree	Betweenness	Eigenvector centrality	Closeness centrality	Ratio Betw./Degree
39	FBLN5	2	0.1194	0.2389	0.2440	0.0597
40	TTLL5	2	0.1182	0.2365	0.2391	0.0591
41	NEK2	2	0.0048	0.0097	0.1766	0.0024
42	WDR19	2	0.0	0.0	0.1667	0.0
43	CTNNA1	2	0.0141	0.0282	0.1081	0.0070
44	MYO7A	2	0.0278	0.0555	0.1205	0.0139
45	ARL3	2	0.0141	0.0282	0.1279	0.0070
46	HK1	2	0.0539	0.1078	0.1678	0.0270
47	FAM161A	2	0.0141	0.0282	0.1736	0.0070
48	USH1C	2	0.0410	0.0821	0.1358	0.0205
49	INPP5E	2	0.2504	0.2504	0.2320	0.1252
50	ACO2	2	0.0141	0.0282	0.1632	0.0070

Table 2: Main network metrics of disease LCC genes.

n	Precision	Recall	F1-score
50	0.024 ± 0.008	0.083 ± 0.025	0.037 ± 0.012
10%	0.114 ± 0.107	0.054 ± 0.050	0.123 ± 0.042
25%	0.067 ± 0.022	0.083 ± 0.025	0.074 ± 0.024
50%	0.033 ± 0.011	0.083 ± 0.025	0.048 ± 0.015
100%	0.019 ± 0.011	0.096 ± 0.052	0.032 ± 0.018

Table 3: Results of 5-fold cross-validation for DIAMOnD.

n	Precision	Recall	F1-score
50	0.032 ± 0.016	0.109 ± 0.051	0.049 ± 0.024
10%	0.143 ± 0.128	0.068 ± 0.059	0.153 ± 0.041
25%	0.089 ± 0.044	0.109 ± 0.051	0.098 ± 0.048
50%	0.044 ± 0.022	0.109 ± 0.051	0.063 ± 0.031
100%	0.025 ± 0.014	0.123 ± 0.063	0.042 ± 0.022

Table 4: Results of 5-fold cross-validation for DiaBLE.

n	Diffusion time	Precision	Recall	F1-score
50	0.002	0.052 ± 0.016	0.179 ± 0.049	0.081 ± 0.024
50	0.005	0.052 ± 0.016	0.179 ± 0.049	0.081 ± 0.024
50	0.010	0.052 ± 0.016	0.179 ± 0.049	0.081 ± 0.024
10%	0.002	0.229 ± 0.069	0.110 ± 0.032	0.149 ± 0.044
10%	0.005	0.229 ± 0.069	0.110 ± 0.032	0.149 ± 0.044
10%	0.010	0.229 ± 0.069	0.110 ± 0.032	0.149 ± 0.044
25%	0.002	0.100 ± 0.042	0.124 ± 0.048	0.111 ± 0.045
25%	0.005	0.100 ± 0.042	0.124 ± 0.048	0.111 ± 0.045
25%	0.010	0.100 ± 0.042	0.124 ± 0.048	0.111 ± 0.045
50%	0.002	0.061 ± 0.021	0.151 ± 0.047	0.087 ± 0.029
50%	0.005	0.061 ± 0.021	0.151 ± 0.047	0.087 ± 0.029
50%	0.010	0.061 ± 0.021	0.151 ± 0.047	0.087 ± 0.029
100%	0.002	0.039 ± 0.010	0.193 ± 0.046	0.065 ± 0.017
100%	0.005	0.039 ± 0.010	0.193 ± 0.046	0.065 ± 0.017
100%	0.010	0.039 ± 0.010	0.193 ± 0.046	0.065 ± 0.017

Table 5: Results of 5-fold cross-validation for heat diffusion.

Figures

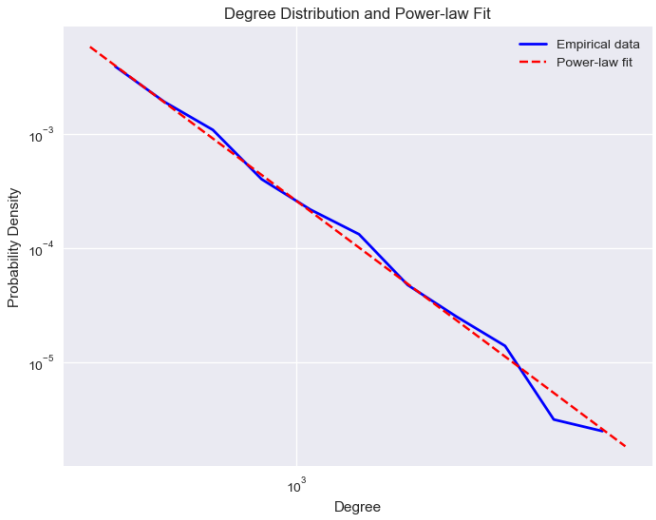

Figure 1: Degree distribution and power-law fit; both axes are in log scale.

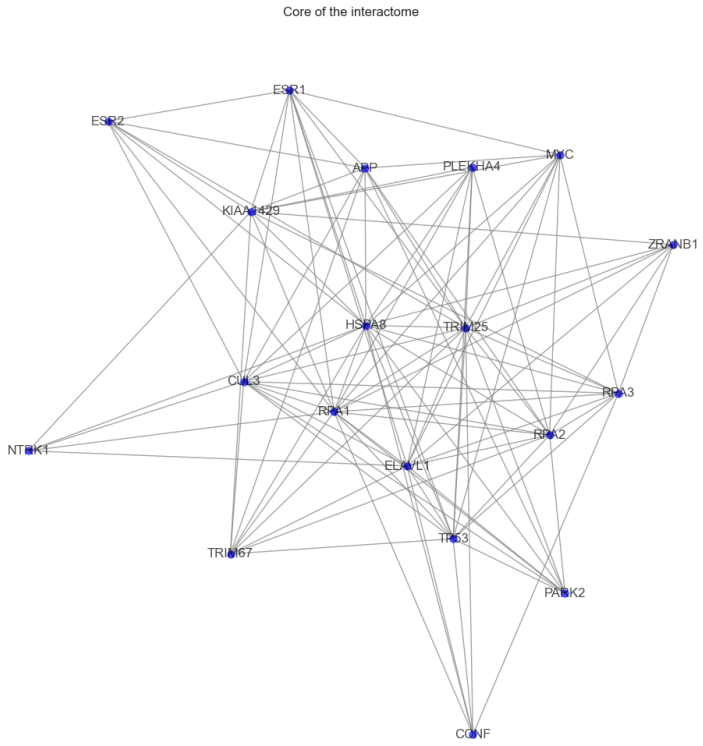

Figure 2: Plot of the network core (subgraph of nodes with degree ≥ 2000).

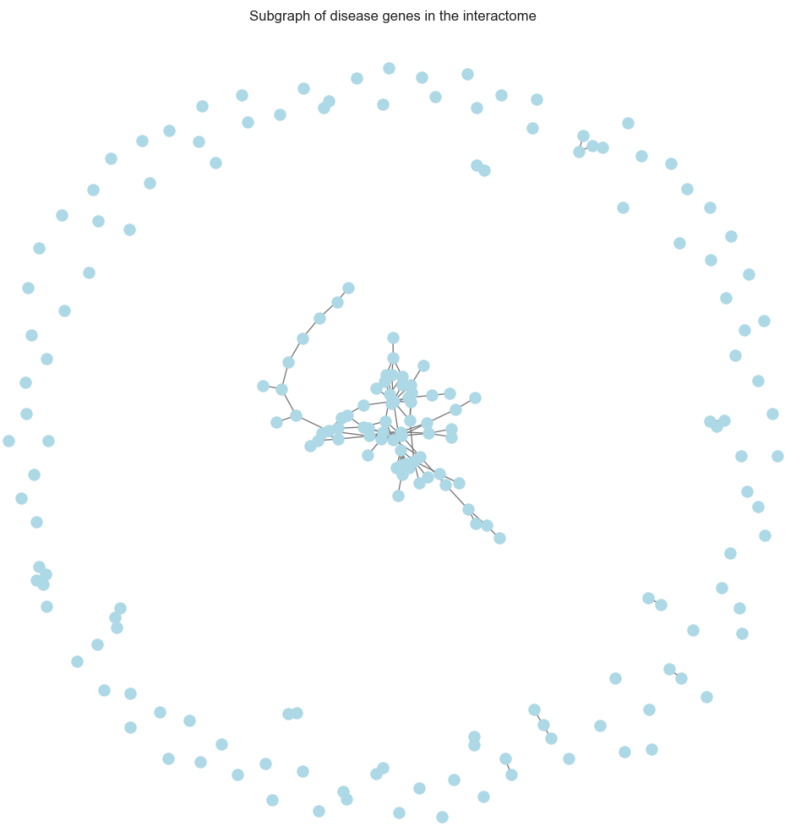

Figure 3: Plot of the disease interactome. Node labels are omitted for visual clarity.

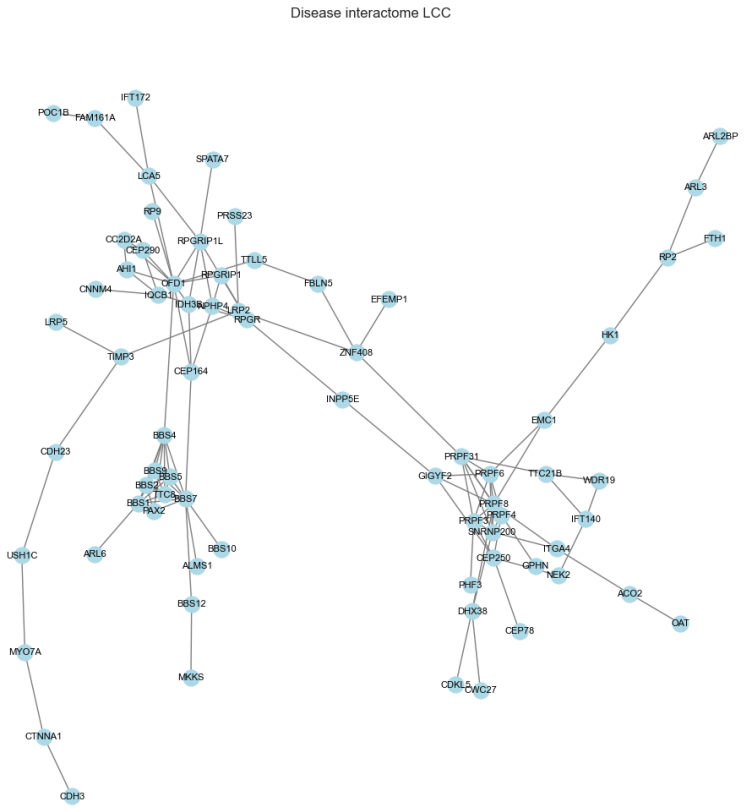

Figure 4: Plot of the LCC of the disease interactome.

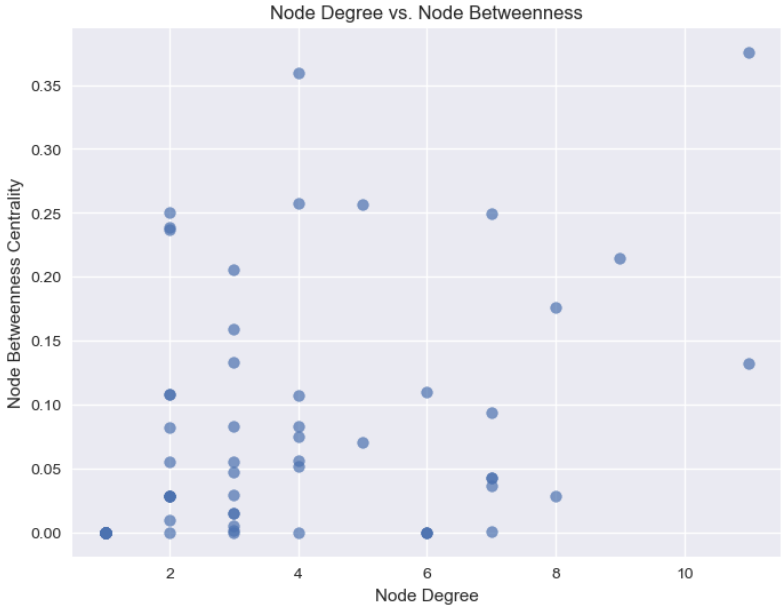

Figure 5: Scatterplot of node degree vs. node betweenness of the disease interactome LCC.

Figure 6: Plot of GDA seed genes and putative disease genes in the interactome. Node labels are omitted for visual clarity.

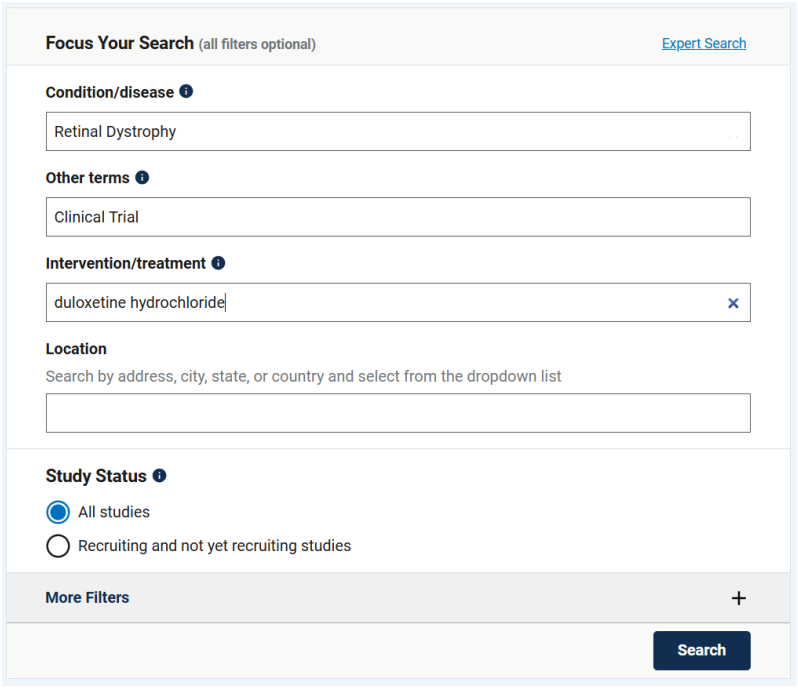

Figure 7: Example of query performed on ClinicalTrials.gov