Inverse Comorbidities: A Networks-Based Analysis

Aadit Mahajan * **Shreya Rajagopalan** *

* *Department of Biotechnology, Indian Institute of Technology Madras.*

Abstract

Inverse comorbidities refer to a lower than expected likelihood of acquiring a certain disease in individuals suffering from a different medical condition. For example, patients with Alzheimer's disease have a lower risk of Lung Cancer. This phenomenon is particularly intriguing, as it can help us gain insight into the pathogenesis of certain diseases and allow us to understand the underlying mechanisms of certain unknown disease pathways. This project aims to find various such inverse comorbidity relations that are manifested due to the onset of a few diseases. Considering the two broad categories of genetic and functional inverse comorbidities, we have performed centrality analysis and community detection on disease-gene association and disease-function association bipartite networks to deduce inverse comorbidity pairs and interesting relations have been uncovered.

Keywords: inverse comorbidities, disease networks, bipartite graphs, louvain algorithm, greedy modularity maximisation, vertex strength

1. INTRODUCTION

Comorbidities are a common phenomenon in many diseases. It refers to the coexistence of more than one disorder or clinical condition in a patient Valderas et al. [2009]. Direct comorbidities indicate a higher than expected co-occurrence of diseases. For example, Alzheimer's disease (AD) and glioblastoma are suspected to be a pair of direct comorbidities Lehrer [2010]. More recently, in the COVID -19 pandemic, diabetes was identified as one of the most common direct comorbidities in affected patients Wu et al. [2020].

However, inverse comorbidities refer to a lower than expected likelihood of acquiring a certain disease in individuals suffering from a different medical condition. For example, patients with AD have a lower risk of Lung Cancer (LC) Driver et al. [2012]. Another well-studied pair is that of sickle-cell anemia and malaria, where sickle cell anemia patients have a lesser than expected probability of acquiring the malarial parasite due to faulty hemoglobin in their red blood cells.

Previously, different experimental methods have been adopted to recognise inverse comorbidities. Transcriptomic meta analyses of gene expression data have been carried out in different disorders to confirm certain inverse comorbidities and to understand causal molecular pathways and relationships Sánchez-Valle et al. [2017],Ibáñez et al. [2014] . A second method is conducting population-based studies to understand the patterns of inverse comorbidity relations Oh et al. [2023].

1.1 Objectives

The number of inverse comorbidities established through the above-mentioned practices currently available in literature is fairly limited. Moreover, a large proportion of these studies is entirely dedicated to cancers and neurodegenerative diseases. This study aims to develop a networks-based model to generalise this and identify a larger set of inverse comorbidity relations that are manifested due to the onset of a few diseases. This will provide valuable insights into the pathogenesis of certain diseases and improve our understanding of the underlying mechanisms of certain unknown disease pathways.

1.2 Why Networks?

Most biological systems are innately complex, both in terms of their molecular components and the interactions between them. In higher organisms such as humans, this can be viewed as networks - intracellular and intercellular, linking tissues and organ systems. A disease or disorder can then be considered as a perturbation in these highly interconnected networks, both structurally and functionally. This suggests that disease is almost always the result of multiple genes or phenotypes in action Barabási et al. [2011]

In terms of modularity, cellular components forming a topological module are functionally closely related, and thus also form a functional module, and since a disease indicates breakdown of a functional module, the functional module also corresponds to a disease module. Thus, disease modules, functional modules, and gene modules are all correlated. Moreover, disease modules display some unique characteristics. A disease module is more likely to overlap with a topological or functional module rather than being

identical in structure with them. Disease modules are defined for a certain disease. Consequently, every disease has a unique disease module. One gene, metabolite or protein can be a part of multiple disease modules, indicating that multiple disease modules can be overlapping with each other Ravasz et al. [2002]. All these characteristics are crucial in identifying disease modules, which would help to uncover interesting inverse comorbodity relations.

The first attempt to create a network of genetically interrelated diseases resulted in the Human Disease Network (HDN) Goh et al. [2007]. Starting with a 'diseasome' bipartite graph with the two sets of nodes as disease phenomes and disease genomes and edges introduced if the mutation of a certain gene has implications on a certain disease, the HDN was generated as a disease projection of this bipartite graph.

Another method of correlation between diseases was developed on the basis of their observed comorbidity, resulting in a Phenotypic Disease Network (PDN) obtained from large number of medical records of patient disease history. Although the PDN does not give insights into the underlying molecular-level mechanisms, it captures information on disease progression as it is observed that patients develop disorders in the network close to the ones that they already carry Hidalgo et al. [2009].

Garnering inspiration from the above formulations of disease networks, we have performed network perturbation and community detection analysis on existing disease-gene and disease-function bipartite network datasets to identify inverse comorbidity pairs of diseases.

2. METHODS

2.1 ***Data***

The primary datasets for the study, the disease-gene (DG) association bipartite network and the gene-function (GF) association bipartite network, were acquired from *Stanford Biomedical Network Dataset Collection* Marinka Zitnik and Leskovec [2018]. In the former, named *DG-AssocMiner* network, genes and diseases are represented by nodes, and the associations between them are represented by edges [Fig. 1]. It contains 519 disease nodes, 7294 gene nodes, and 21357 edges. In the latter, named *GF-Miner* network, genes and their Gene Ontologies (GO) or biological functions are represented by nodes, and the functional annotations between them are represented by edges. This contains 16628 function nodes, 5957 gene nodes, and 16628 edges.

2.2 ***Softwares***

Python3 was used to perform the coding tasks for this study. In particular, packages like *NetworkX* for computing graph properties, *Numpy* for computation, *Pandas* for data management, *Scikit-Network* for community detection, *Matplotlib* for data visualisation, and *Cytoscape* for graph visualisation were used.

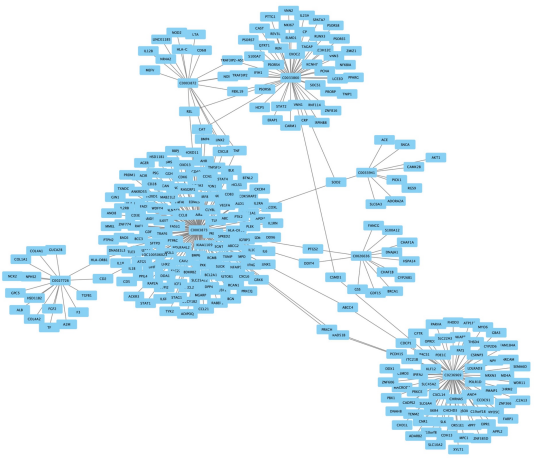

Figure 1. Visualisation of a small subset of the Disease-Gene Association Network

2.3 ***Workflow***

The search for inverse comorbodity pairs is split into two branches since these diseases can be related to each other based on two major classes - genetic and functional. In this study we focus on inverse comorbodities of genetic origin and the workflow is designed according to this goal.

2.3.1 Pruning the search space

The DG network is a fairly large dataset to work with on personal computers. Moreover, it contains certain genes that are ubiquitous, meaning they are expressed in a large fraction of pathways, and mutations in these genes would be causal to a large number of diseases. Consequently, network perturbations involving these gene deletions are more likely to cause fatal changes to the system than others. Hence, the network has been pruned by ignoring these genes in further analyses as they would not result in significant inferences. This has been done by utilising relations from the GF network dataset. The two datasets were appropriately calibrated to account for differences in representation. Genes in the GF network having a degree centrality higher than a threshold of the 90th percentile of degree centralities, i.e., affecting a large number of functional pathways, are ignored while working with the DG network. This resulted in a significant reduction of the number of edges in the DG network to 8043.

2.3.2 Disease projection network

Since our objective is to infer correlations between the different diseases, we restricted our workspace to a weighted disease projection of the pruned bipartite network. This also significantly reduced computational cost and time taken to run the code. The disease projection network contains nodes representing diseases and an edge present between two diseases if there is at least one gene connecting the two diseases in the original *DG-AssocMiner* network. Edge weights are assigned to capture the number of genes that link two particular diseases in the original bipartite graph.

2.3.3 Weights and Centralities

The degree centrality of a vertex is defined as

$$k_i = \sum_j a_{ij} \tag{1}$$

where a_{ij} refers to the elements of the adjacency matrix.

Eigenvector centrality of a vertex is defined as

$$e_{ij} = \alpha \sum_j a_{ij} e_j \tag{2}$$

where e_i denotes the eigenvector centrality of node i, a_{ij} denotes the elements of the adjacency matrix and α is some constant.

The above two measures of centralities were calculated for the vertices of the disease projection network and sorted accordingly to derive insights into the most important nodes. However, these are purely node attributes and do not take edge weights into consideration, thus resulting in a loss of information corresponding to gene relations.

A more useful node attribute, the vertex strength s_i, is calculated by incorporating edge weights into the vertex degree Barrat et al. [2004], i.e., combining information about disease and gene relations. It is defined as :

$$s_i = \sum_{j=1}^{N} a_{ij} w_{ij} \tag{3}$$

where a_{ij} refers to the elements of the adjacency matrix (a measure of degree) and w_{ij} refers to the edge weight of the edge connecting the i^{th} and j^{th} nodes and N is the total number of nodes.

2.3.4 Modularity and Community detection

Modularity of a network Q is defined as

$$Q = \frac{1}{(2m)} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{(2m)} \right] \delta(c_i, c_j) \tag{4}$$

where c_i and c_j refer to the groups that the i^{th} and j^{th} nodes belong to, respectively, and δ_{xy} denotes the Kronecker delta function, m is the number of edges, A is the adjacency matrix, k_i and k_j are the degrees of the i^{th} and j^{th} nodes respectively.

Modularity is a particularly favourite metric for community detection as it records the density of links between and within communities, thus giving a measure of how strong the resulting communities are. Hence, maximising this property is a sure-shot way to obtain good community structures.

For this study, community detection was carried out on the weighted disease projection network using the Clauset-Newman-Moore greedy modularity maximization algorithm Clauset et al. [2004] and Louvain Algorithm Blondel et al. [2008] available in the NetworkX package.

2.3.5 Zeroing in on inverse comorbidities

The clusters are first obtained by running the Louvain community detection algorithm on the disease projection network. These clusters are a good representation of diseases that share many common genes. Inside these clusters, disease pairs that have the least edge weights indicate pairs that have significant overlapping between their gene neighbourhood in the DG network.

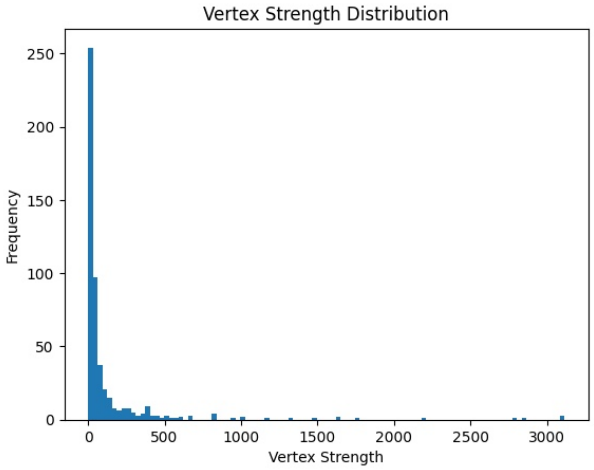

Figure 2. Distribution of vertex strength over all the nodes shows a power law-like relation

Disease	Vertex strength
Mammary Neoplasms	3112
Muscle Weakness	3100
Down Syndrome	3096
Mammary Neoplasms, Experimental	2861
Stomach Neoplasms	2201

Table 1. Few of the important nodes that exhibit high vertex strengths

These pairs were then scrutinized further to check whether there was a significant overlap in the genetic neighbourhoods of the diseases. For this, the neighbourhood of each disease in a pair was calculated and an overlap fraction was obtained for each disease. This overlap fraction determined how significant the overlap was for a disease. A score was calculated using the overlap fractions of the two pairs of diseases. This score was then used to sort all the pairs obtained, and the pairs with the highest scores were analysed and compared to results from literature.

3. RESULTS

3.1 ***Important nodes***

The importance of nodes was measured by their vertex strength. These nodes signify the diseases that have a very large number of gene-associations. Diseases that showed to have high vertex strengths included *Mammary Neoplasms, Down Syndrome, Stomach neoplasms, Systemic arterial pressure, etc* [Table 1]. The network reflects the significance of cancers and their gene-associations, as 12 out of the top 20 highest vertex strength diseases were types of cancer (or neoplasms).

3.2 ***Important edges***

The edges that have high edge weights show high correlation between the two diseases that they connect. As shown in table 2, the correlations are almost natural and

Disease 1	Disease 2	Edge Weight
Muscle Weakness	Down's Syndrome	55
Animal Mammary Neoplasms	Mammary Neoplasms, Experimental.	50
Arsenic Poisoning	Dermatological disorders	24
Adenoid Cystic Carcinoma	Salivary Gland Neoplasms	23
Mammary Neoplasms	Neoplasm Metastasis	19

Table 2. Few of the important edges in the disease projection network

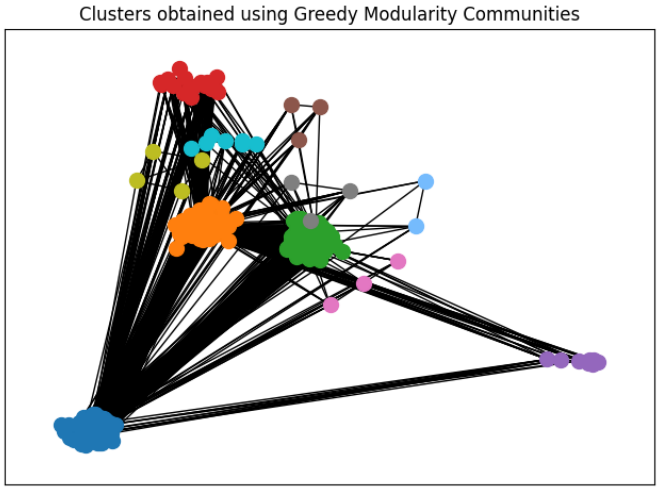

Figure 3. Visualisation of communities. Different coloured nodes correspond to different communities.

predictable. For example, *muscle weakness* is highly linked to *Down's syndrome* and *mammary neoplasms* show a high correlation to *neoplasm metastasis*.

3.3 ***Important communities***

Communities obtained from using greedy modularity community detection are shown in *fig.2* as a multicolored visualisation. The hubs of these communities (along with their colour in *fig.2*) were found to be the diseases *Rheumatoid Arthritis (blue), Mammary Neoplasms (orange), Obesity (green), Leigh Disease(red), Neural Tube defects (purple), Autoimmune Diseases (cyan), Zellweger syndrome (olive), Precursor T-cell Lymphoblastic Leukemia-Lymphoma (pink), Respiratory Hypersensitivity (grey), Meningioma (brown), Romano-Ward Syndrome (sky blue)*. Looking at these communities gives a fair idea of how well-linked each community is within and with other communities. *Mammary Neoplasms* cluster (blue) shows high modularity and extensive linkage to other communities, reinforcing the large number of genes the family of cancers affects. Other significant clusters include the *Rheumatoid arthritis* cluster and the *Obesity cluster*, which are also major players in the disease-gene associations.

Disease 1	Disease 2	Overlap Fraction
Marfan Syndrome	Metabolic Bone Disorder	0.593
Ataxia	Brain Diseases	0.588
Sarcoidosis	Urticaria	0.563
Diabetes Mellitus	Pre-Eclampsia	0.481
Hepatitis B	Sarcoidosis	0.417

Table 3. A compilation of few of the results obtained for inverse comorbidities

3.4 ***Inverse comorbidities***

The disease pairs obtained as a result of the method used for finding are shown in Table 3. The results are of mixed inferences and few of them show traits of inverse comorbidities. For example, *Marfan Syndrome* and *Metabolic bone disorder* have a direct genetic correlation of being a comorbidity. Whereas *Sarcoidosis* and *Urticaria* have very low relation in terms of affected parts of the body, indicating they might be an inverse comorbidity. The same is the case with the fourth example of *Diabetes Mellitus* and *Pre-Eclampsia*, which have a seemingly low correlation in terms of affected tissues/organs.

4. DISCUSSION

4.1 ***Interpretation of results***

The results obtained for all the analyses done using various network analysis approaches show interesting results. The disease-gene association network definitely showed traits that were expected of the network, such as power law-like behaviour. The results obtained from analysing the important edges, nodes, and clusters also walk hand in hand with results obtained in conventional biology and medicine, where we find well-documented disease correlations and disease significance. Finally, the results obtained for inverse comorbidities also look promising, although further pruning and selection of correct pairs are required to find the exact inverse comorbidity pairs.

4.2 ***Assumptions***

A few assumptions were made during the creation of this model to predict inverse comorbidities. Firstly, it was assumed that it was possible to segregate inverse comorbidity relations into two types - functional and genetic inverse comorbidity. This assumption may not be a bad one as diseases, in general, can be classified into two broad categories – pathogen-related or genetic. In this analysis, we have only looked into the genetic aspect of the network data that was available to us. Secondly, we have assumed that the dataset is complete in terms of the disease-gene associations for each mentioned gene. This is an assumption that must be made if the dataset is to be used to derive any results.

4.3 ***Limitations***

This study mostly focused on the genetic aspect of disease association networks, and the perspective can be broadened by extending these ideas to disease-function association networks. Secondly, the results obtained for finding

inverse comorbidities are mixed with certain pairs of comorbidity relations, also indicating that there is scope for improvement to narrow down further on the exact pairs. However, these pairs are not exactly diseases of purely genetic origin, and such pairs should fall into the category of inverse comorbidity. We could not find a dataset of purely genetic disease where we could map only those diseases that had genetic origin as the cause.

5. FUTURE WORK

Taking forward the realisation of the two broad categories of inverse comorbidities - genetic and functional, we have already started with a similar analysis of the disease-function association bipartite network and aim to uncover similarities and differences between the two approaches.

With appropriate datasets, this project can be extended to finding diseases that have been evolutionarily selected in cases of inverse comorbidity (for example, sickle cell anaemia and malaria). Discovering the root evolutionary causes of such diseases may help identify other pathogenic pathways and improve our understanding of pathogenesis.

6. CONCLUSIONS

The method/model proposed to be used to find inverse comorbidities may not be foolproof, although the results obtained were interesting from a clinical as well as a biological perspective.

7. ACKNOWLEDGEMENTS

We would like to thank Prof. Karthik Raman for giving us the wonderful opportunity to perform this project as part of the course BT5240: Computational Systems Biology.

8. SUPPLEMENTARY INFORMATION

All the code used for this project can be found on https://github.com/aadit-mahajan/GetTheseDiseases.

REFERENCES

Barabási, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. doi:10.1038/nrg2918.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101. doi:10.1073/pnas.0400087101.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008. doi:10.1088/1742-5468/2008/10/P10008.

Clauset, A., Newman, M.E., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 70. doi:10.1103/PhysRevE.70.066111.

Driver, J.A., Beiser, A., Au, R., Kreger, B.E., Splansky, G.L., Kurth, T., Kiel, D.P., Lu, K.P., Seshadri, S., and Wolf, P.A. (2012). Inverse association between cancer and alzheimer's disease: Results from the framingham heart study. *BMJ (Online)*, 344. doi:10.1136/bmj.e1442.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabási, A.L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104. doi:10.1073/pnas.0701361104.

Hidalgo, C.A., Blumm, N., Barabási, A.L., and Christakis, N.A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5. doi:10.1371/journal.pcbi.1000353.

Ibáñez, K., Boullosa, C., Tabarés-Seisdedos, R., Baudot, A., and Valencia, A. (2014). Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses. *PLoS Genetics*, 10. doi:10.1371/journal.pgen.1004173.

Lehrer, S. (2010). Glioblastoma and dementia may share a common cause. *Medical Hypotheses*, 75. doi:10.1016/j.mehy.2010.01.031.

Marinka Zitnik, Rok Sosič, S.M. and Leskovec, J. (2018). BioSNAP Datasets: Stanford biomedical network dataset collection. http://snap.stanford.edu/biodata.

Oh, J., Lee, H.S., Jeon, S., Seok, J.H., Yoo, T.K., Park, W.C., and Yoon, C.I. (2023). Marked reduction in the risk of dementia in patients with breast cancer: A nationwide population-based cohort study. *Cancer Research and Treatment*, 55. doi:10.4143/crt.2022.272.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297. doi:10.1126/science.1073374.

Sánchez-Valle, J., Tejero, H., Ibáñez, K., Portero, J.L., Krallinger, M., Al-Shahrour, F., Tabarés-Seisdedos, R., Baudot, A., and Valencia, A. (2017). A molecular hypothesis to explain direct and inverse co-morbidities between alzheimer's disease, glioblastoma and lung cancer. *Scientific Reports*, 7. doi:10.1038/s41598-017-04400-6.

Valderas, J.M., Starfield, B., Sibbald, B., Salisbury, C., and Roland, M. (2009). Defining comorbidity: Implications for understanding health and health services. *Annals of Family Medicine*, 7. doi:10.1370/afm.983.

Wu, J., Zhang, J., Sun, X., Wang, L., Xu, Y., Zhang, Y., Liu, X., and Dong, C. (2020). Influence of diabetes mellitus on the severity and fatality of sars-cov-2 (covid-19) infection. *Diabetes, Obesity and Metabolism*, 22. doi:10.1111/dom.14105.