Master Degree: Data Science 15 January 2025
a.y. 2024/2025

BIOINFORMATICS AND NETWORK MEDICINE

Putative disease gene identification and drug repurposing for Osteoporosis

Syed H. Bashar, Emre Pelzer, Ayesegul S. Ozgenkan

GROUP 14

ABSTRACT

This study identifies therapeutic targets and drug candidates for osteoporosis using network-based methods. The human interactome was reconstructed from BioGRID data, and disease-related genes were extracted from curated gene-disease associations. Three algorithms—DIAMOnD, DiaBLE, and Diffusion-based—were used to infer putative disease genes, with the best-performing algorithm (DIAMOnD) predicting 100 new candidates. Enrichment analyses and drug repurposing identified potential pathways and drugs, with clinical trials validation conducted for the top-ranked candidates. This approach provides insights into osteoporosis therapeutics.

INTRODUCTION

Osteoporosis, a progressive bone disease, leads to increased fragility and fracture risk, significantly impacting the quality of life in affected individuals. Current therapies primarily focus on managing symptoms and slowing bone loss, but a deeper understanding of the disease's molecular underpinnings is essential for developing targeted treatments. Advances in computational biology and network-based analyses offer powerful tools for exploring the intricate interactions between genes and proteins involved in disease processes.

In this study, we leverage network biology to investigate osteoporosis by reconstructing the human interactome and identifying disease-related genes from curated datasets. Using a comparative approach, we employ multiple algorithms to infer putative disease genes and validate their relevance through enrichment analyses. Further, we explore drug repurposing opportunities by linking prioritized genes to approved drugs, highlighting promising candidates for clinical validation. This comprehensive framework aims to bridge the gap between molecular research and therapeutic applications in osteoporosis.

MATERIALS AND METHODS

Describe in this section the experimental procedures and resources, data analysis procedures and statistical methods. Give enough detail to replicate the experiment but do not overwhelm the reader with too many details. For what concerns the specific work, please follow these main steps:

1. PPI and GDA data gathering and interactome reconstruction

Protein-Protein Interaction (PPI) Data: To construct the human interactome, the latest "all organisms" tab3 file was downloaded from the BioGRID database. The dataset was filtered to retain only interactions where both "organism A" and "organism B" were equal to 9606 (Homo sapiens). Interactions labeled as "physical" (Experimental System Type = physical) were selected, while redundant entries and self-loops were removed to ensure data quality. The largest connected component (LCC) of the human interactome was then isolated, representing the largest subset of connected nodes and edges.

Gene-Disease Association (GDA) Data: Osteoporosis-related genes were obtained from the curated GDA file DISEASES_Summary_GDA_CURATED_osteoporoses.tsv. The "Gene" column was extracted, and gene names were verified against the HGNC database for accuracy. Identifiers were cross-checked using columns such as UnitProt, geneEnsemblIDs, geneNcbiID, and geneNcbiType, and any inconsistencies were resolved to ensure uniformity and correctness. Building the Disease LCC: The disease-associated genes identified in the previous step were mapped onto the interactome LCC to determine their presence. Interactions involving only the disease-associated genes were extracted to construct the disease interactome. The largest connected component (LCC) of this disease interactome was then isolated for further analysis.

Table 1: Summary of GDAs and basic network data

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*
Osteoporosis	C0029456	T047	85	84	47

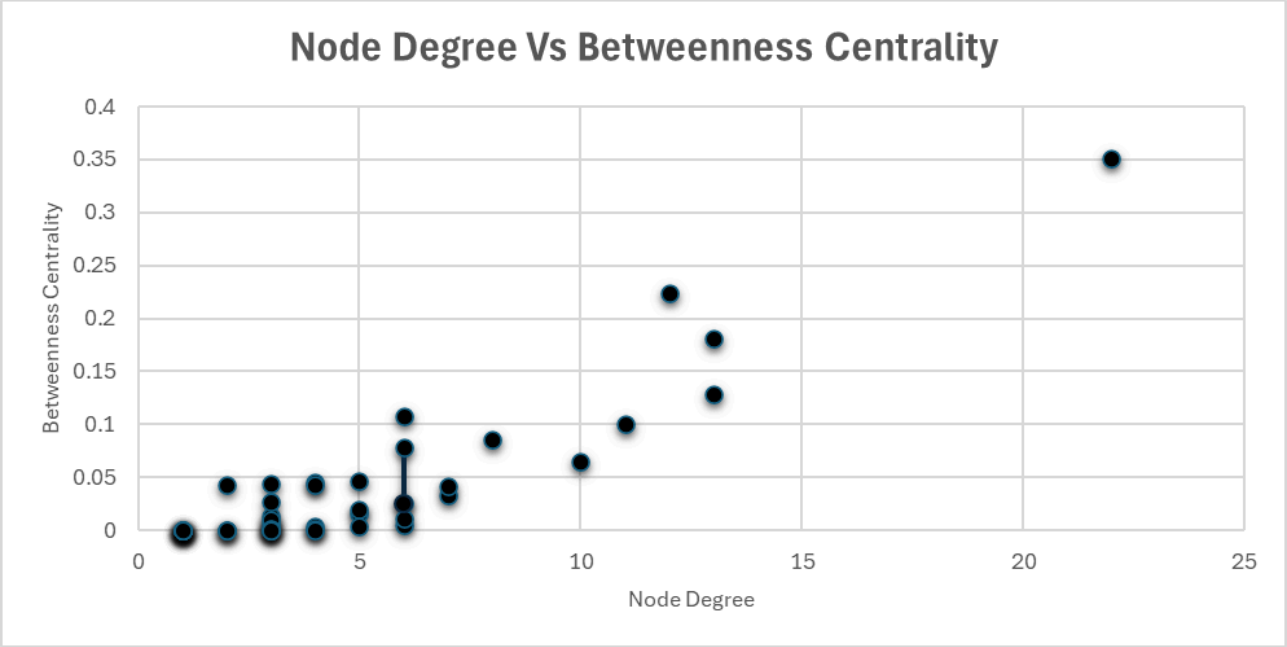
Network Metrics and Characterization: Key network metrics were computed for all nodes (genes) in the disease LCC: Node Degree, Betweenness Centrality, Eigenvector Centrality & Closeness Centrality. The network metrics were tabulated for the top 10 disease-associated genes in the disease LCC, ranked by node degree, as shown in Table 2. A scatterplot (Fig:1) was generated to visualize the relationship between node degree and betweenness centrality, offering a clearer understanding of the network's structure and key nodes within the disease interactome.

Table 2: Main network metrics of disease LCC genes(1st 10 rows)

Node	Node Degree	Betweenness Centrality	Eigenvector Centrality	Closeness Centrality	Betweenness/ Degree
ESR2	22	0.35063606	0.465064144	0.567901235	0.015938003
ESR1	13	0.128507987	0.304439573	0.484210526	0.00988523
SRC	13	0.181351963	0.31497338	0.5	0.013950151
ENO1	12	0.224413515	0.256061606	0.522727273	0.018701126
GAPDH	11	0.100957395	0.237012318	0.479166667	0.009177945
PARK7	10	0.065250734	0.137030573	0.4	0.006525073
TPI1	8	0.086162662	0.153290436	0.425925926	0.010770333
CCT2	7	0.032903203	0.191722841	0.442307692	0.004700458
ACTG1	7	0.042035072	0.169749373	0.438095238	0.00600501

CAP1	6	0.107311344	0.14110134	0.433962264	0.017885224

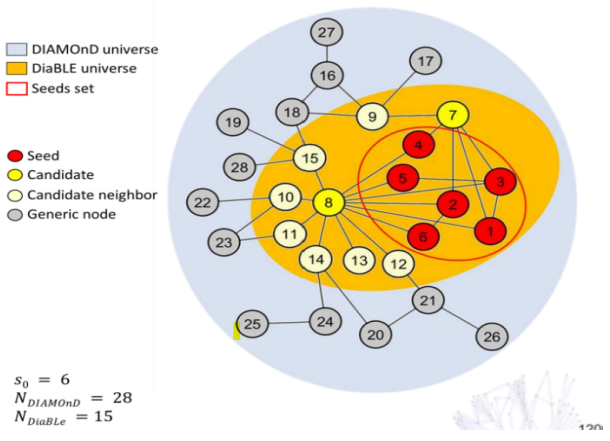
Fig1: Node Degree vs Betweenness Centrality



2. Comparative analysis of the disease genes identification algorithms

A 5-fold cross validation was performed by spitting the disease genes in sets of 5 subsets and setting one as the probe set S_P . Each time, the remaining 4 sets were used as training set S_T and the other was used for testing. three different algorithms were used:

- **DIAMOnD algorithm**: The DIAMOnD algorithm identifies candidate disease genes by analyzing their network proximity to known disease-associated genes within a protein-protein interaction (PPI) network. Starting with a set of known seed genes, the algorithm scores all nodes (genes) in the network based on their connectivity to the seed genes using a hypergeometric test to assess the statistical significance of their interactions. Nodes are then ranked by their scores, and the most significant node is iteratively added to the disease module. This process continues until the desired number of candidate genes is identified. The final output is a prioritized list of genes that are likely to be associated with the disease, based on their network-level interactions with the seed genes.



- **DiaBLE algorithm**:This is a modified version of the DIAMOnD algorithm where the universe size is modified by taking the cluster of the disease genes and expanding the universe to include the nodes connected to the disease genes and first neighbors.

- **Diffusion-based algorithm**:The Diffusion-based algorithm identifies candidate disease genes by simulating the spread of "heat" or information across a protein-protein interaction (PPI) network over specified diffusion times. Starting with a set of known seed genes (assigned an initial "heat" value of 1), the algorithm models the diffusion process across the network. Nodes closer to the seed genes

accumulate higher heat values, while those further away receive less heat as the diffusion progresses. The algorithm is run for three different diffusion times: t=0.002, t=0.005, and t=0.01. At the end of each diffusion process, nodes are ranked based on their heat values, with higher values indicating stronger network proximity to the seed genes. The final output is a ranked list of candidate genes likely to be associated with the disease, identified by their connectivity and heat distribution in the network.

All three of the algorithms we run and the performance metrics for them were calculated. Here the performance metrics considered were precision, recall and F1-score. The obtained results are shown in table 3. From the results obtained, the best performing algorithm was DIAMOnD.

Table 3: performance metrics for the three algorithms used.

Algorithm	Precision (mean ± SD)	Recall (mean ± SD)	F1-Score (mean ± SD)
DIAMOnD	0.0200 ± 0.0126	0.0625 ± 0.0395	0.0303 ± 0.0192
DiaBLE	0.0160 ± 0.0196	0.0500 ± 0.0612	0.0242 ± 0.0297
Diffusion_t=0.002	0.0080 ± 0.0098	0.0250 ± 0.0306	0.0121 ± 0.0148
Diffusion_t=0.005	0.0080 ± 0.0098	0.0250 ± 0.0306	0.0121 ± 0.0148
Diffusion_t=0.01	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000

3. Putative disease gene identification

The best performer, DIAMOnD algorithm was used to create a list of 100 putative genes using the script provided here. To run the script the edge list of the Interactome_LCC and the 84 disease genes found in the LCC were used.

Enrichment analysis: Enrichment analysis was performed using the 100 putative gene and the original disease gene sets. The comparison contained GO-BP, GO-MF, GO-CC, Reactome and KEGG pathways. The results of the analysis are shown in table 4. The gene names were put in EnrichR to obtain the results. Only adjusted p-value of < 0.05 was considered.

Table 4: Obtained data from EnrichR (adj. p-value < 0.05)

Category	Total Original Terms	Total Putative Terms	Number of Overlapping Terms	Percentage Overlap (Original)	Percentage Overlap (Putative)	Unique Terms in Original	Unique Terms in Putative
GO-BP	387	133	29	7.49	21.80	358	104
GO-MF	27	36	5	18.52	13.89	22	31
GO-CC	22	16	9	40.91	56.25	13	7

Reactome	240	174	72	30.00	41.38	168	102
KEGG	74	28	12	16.22	42.86	62	16

The results show varying degrees of overlap between the original and putative gene sets across five functional categories. **GO-CC** exhibited the highest overlap (40.91% original, 56.25% putative), indicating strong agreement in cellular localization. **Reactome pathways** also showed substantial overlap with 72 shared terms (30.00% original, 41.38% putative). In contrast, **GO-BP** had a lower overlap (7.49% original, 21.80% putative), highlighting significant functional diversity. **GO-MF** and **KEGG pathways** had moderate overlaps, with putative genes contributing additional molecular functions and pathways. Overall, the putative genes captured key pathways and biological functions while introducing novel terms for further investigation.

4. Drug repurposing

Drug identification: The first 20 putative disease genes (lowest p-value) were collected and using the latest "interactions.tsv" from DGIdb a list of approved drugs for the disease genes were collected. out of 20, only 8 had associated drugs that were approved. CDK2 had the highest number of drugs associated (9).

Clinical Trials validation: The first three drugs of each gene were individually checked here for clinical trials testing. Out of the eight genes only NME2 had two drugs LAMIVUDINE(1) & TENOFOVIR(3) in clinical trials.

Table 5: genes with their associated drugs list

gene_name	associated_drugs	drug_count
CDK2	['LOVASTATIN', 'PACLITAXEL', 'RESVERATROL', 'DAUNORUBICIN LIPOSOMAL', 'CARBOPLATIN', 'ERIBULIN MESYLATE', 'ACETAMINOPHEN', 'RALTITREXED', 'DEXAMETHASONE']	9
BAP1	['PANOBINOSTAT', 'VORINOSTAT', 'SUNITINIB', 'EVEROLIMUS', 'VALPROIC ACID', 'OLAPARIB']	6
CFL1	['CLOTRIMAZOLE', 'CINNARIZINE', 'CLOFIBRATE', 'FENOFIBRATE MICRONIZED', 'SERTRALINE HYDROCHLORIDE']	5
NME2	['LAMIVUDINE', 'TENOFOVIR', 'ZIDOVUDINE', 'PROGESTERONE', 'ADEFOVIR DIPIVOXIL']	5
SOD1	['TETRACYCLINE', 'DOXYCYCLINE ANHYDROUS', 'TOFERSEN', 'OXYTETRACYCLINE ANHYDROUS']	4
EZH2	['TAZEMETOSTAT', 'DABRAFENIB', 'TAZEMETOSTAT HYDROBROMIDE']	3
FN1	['OCRIPLASMIN', 'DACARBAZINE']	2
VCP	['HEXACHLOROPHENE']	1

5. ProConSuL Vs DIAMOnD

ProConSuL is an algorithm designed for prioritizing disease-associated genes based on their network topology and disease relevance. When comparing the top 20 prioritized genes identified by ProConSuL and DIAMOnD, a significant overlap of 15 genes was observed, highlighting a shared focus on key disease-related genes between the two approaches. These overlapping genes include notable ones like AGR2, CLIC4, BAP1, and CDK2, which are likely central to the disease network. ProConSuL uniquely identified 5 genes highlighted in the table in red. The same is done for five unique genes identified by the DIAMOnD algorithm. This comparison underscores both the shared and distinct strengths of each algorithm in capturing different aspects of the disease network.

ProConSul_node	DIAMOnD_node
AGR2	AGR2
BAP1	TIMP2
VCP	CDK2
CDK2	CLIC4
CLIC4	PRDX6
PRDX6	CLIC1
CLIC1	BAP1
HSPA8	VCP
EZH2	DSTN
UBE2M	MYOC
DSTN	FN1
PRDX2	TKT
SOD1	CFL1
CFL1	UBE2M
IQGAP1	SOD1
ISG15	PRDX2
U2AF2	NME2
ACO2	LDHA
MCM2	EZH2
TKT	ACO2

RESULTS AND DISCUSSION

The reconstruction of the human interactome, combined with disease gene mapping for osteoporosis, enabled the identification of key disease-related genes and their interactions within the largest connected component (LCC). Network analysis highlighted the centrality and connectivity of specific nodes, suggesting their importance in the disease network. Comparative analysis of gene identification algorithms demonstrated varying levels of performance, with DIAMOnD emerging as the most effective approach for predicting putative disease genes. This analysis emphasized the ability of network-based methods to prioritize candidate genes based on their proximity to known disease-associated genes.

The putative disease genes were further analyzed for their therapeutic relevance by linking them to approved drugs using DGIdb. Several genes showed multiple drug associations, with some drugs standing out as promising candidates for further investigation. Clinical trials validation revealed that only a few drugs, such as **Tenofovir** and **Lamivudine**, had ongoing trials related to osteoporosis, while others lacked clinical evidence. These findings demonstrate the potential of integrating network analysis, gene prioritization algorithms, and drug repurposing strategies to identify novel therapeutic opportunities and guide future research for osteoporosis treatment.

AUTHOR CONTRIBUTIONS

For the project work distribution, Syed Habibul Bashar was responsible for PPI and GDA data gathering and Interactome reconstruction. He was also responsible for the comparative analysis of the disease genes identification algorithms. Emre Pelzer was responsible for putative disease gene identification and drug repurposing. Aysegul Sine Ozgenkan helped with the clinical trial validation and the running the PROCONSUL algorithm and comparing the results with the best performer algorithm from part 2 of the report.

REFERENCES

1. Oughtred, R., Stark, C., Breitkreutz, B. J., Rust, J., Boucher, L., Chang, C., ... & Tyers, M. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research, 47*(D1), D529-D541. https://doi.org/10.1093/nar/gky1079

2. Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., ... & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics, 14*(128). https://doi.org/10.1186/1471-2105-14-128
3. Freshour, S. L., Kiwala, S., Cotto, K. C., Coffman, A. C., McMichael, J. F., Song, J. J., ... & Griffith, M. (2021). Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Research, 49*(D1), D1144-D1151.
4. Ghiassian, D. (n.d.). *DIAMOnD: Disease Module Detection*. GitHub repository. Retrieved from https://github.com/dinaghiassian/DIAMOnD
5. Deluca, R. (n.d.). *PROCONSUL*. GitHub repository. Retrieved from https://github.com/rickydeluca/PROCONSUL