Master Degree: Data Science
Academic Year : 2023/2024

5 November 24

BIOINFORMATICS AND NETWORK MEDICINE

Putative disease gene identification and drug repurposing for Obesity

Rahim Rahimov

ABSTRACT

In this study, I investigated multiple methodologies to analyze protein-protein interactions (PPIs) and their significance in disease-gene associations, specifically targeting Obesity. Obesity is a chronic, multifactorial disease characterized by an excessive accumulation of adipose tissue and often associated with significant metabolic dysregulation, including insulin resistance, inflammation, and hormonal imbalances. The objective was to identify the most effective algorithm for accurately extracting genes linked to obesity-related pathways. The algorithms evaluated included DIAMOnD (for disease module detection), DiaBLE, and a heat diffusion approach. Comparative performance was assessed using 5-fold cross-validation and evaluation metrics such as precision, recall, and F1 score. Results indicate that DiaBLE, combined with enrichment analysis, provided the most robust gene identification. Additionally, drug repurposing was explored to identify existing drugs that could target the specific genes implicated in obesity, potentially offering new therapeutic options. All analyses were conducted using Python.

INTRODUCTION

Obesity is a complex global health issue linked to numerous metabolic and cardiovascular diseases, driven by genetic, environmental, and behavioral factors. Protein-protein interaction (PPI) networks offer insights into the cellular interactions underlying obesity, highlighting functional relationships among proteins involved in lipid metabolism, inflammation, and adipogenesis. By mapping these networks, researchers can pinpoint critical "hub" proteins and pathways that may serve as biomarkers or therapeutic targets, paving the way for precision medicine in obesity management [1]

PPI network analysis also aids in identifying potential intervention points within obesity-related pathways, helping to develop targeted therapies that address the disease's molecular roots. Integrating PPI analysis with genomic data thus accelerates the discovery of biomarkers and drug targets, offering promising directions for personalized treatment and prevention strategies [2].

MATERIALS AND METHODS

PPI and GDA data gathering and interactome reconstruction

Protein-protein interaction (PPI) networks are mathematical constructs representing the physical interactions among proteins within the cell. In this study, the Human Protein-Protein Interaction network was constructed using data from BioGRID (The Biological General Repository for Interaction Datasets). Starting with the raw dataset, i filtered out all non-human interactions and retained only “physical” interactions, removing redundant self-loops to ensure a streamlined network structure. This resulted in a network comprising 19,936 nodes and 766,017 edges. The largest connected component of this network, containing all 19,936 nodes, was then isolated for further analysis.

To identify seed genes associated with obesity, i utilized DisGeNET, a comprehensive repository of human gene-disease associations (GDAs). A total of 288 obesity-related genes were mapped within the human interactome. In the largest connected component of the disease-specific network, i calculated several centrality metrics, including Node Degree, Betweenness Centrality, Eigenvector Centrality, Closeness Centrality, and the Betweenness-to-Degree ratio. Table 2 presents the top 50 disease-associated genes in this component, ranked by the Betweenness-to-Degree ratio in descending order

Putative disease genes identification algorithms

In this section, i employed various algorithms to identify putative disease genes, recognizing that disease genes do not distribute randomly within protein-protein interaction (PPI) networks but tend to cluster in specific regions, forming distinct disease modules. Identifying these modules is crucial for understanding the biological mechanisms underlying diseases and for pinpointing potential drug targets. The DIAMOnD algorithm identifies genes connected to seed genes based on significant connectivity, providing valuable network context for our set of putative obesity-associated genes and facilitating the expansion of the gene set in a methodologically sound manner.

DiaBLE, a modification of DIAMOnD, differs by defining a dynamic gene universe: rather than using the entire interactome as a background model, it iteratively limits the gene universe to the smallest local expansion around the current seed set. The third algorithm, Heat Diffusion, operates on network propagation principles to approximate distances within gene networks. This approach leverages PPI networks to highlight new genes relevant to established sets of disease-associated genes. By examining these networks, we hypothesize that genes interacting with known disease genes are likely to contribute to the disease phenotype.

To validate these algorithms' performance, i applied 5-fold cross-validation, segmenting the disease gene set into five subsets, iteratively using one subset as a test set and the remaining four for training.

Best algorithm choice and putative disease gene identification

Using the best-performing algorithm, i aimed to predict new putative disease genes by employing all known Gene-Disease Associations (GDAs) as seed genes. Enrichment Analysis was then conducted on the top 100 genes identified, utilizing the Enrichr platform. This method allows for the identification of gene or protein classes over-represented within a large dataset, potentially linked to disease phenotypes. Statistical techniques are used to pinpoint significantly enriched gene groups, enabling us to detect specific Gene Ontology (GO) categories or biological pathways that are enriched beyond random expectation.

The first technique employed, Gene Ontology (GO) Analysis, systematically associates gene collections with functional biological terms, providing insight into which categories are enriched for the given gene list. Additionally, Pathway Analysis was conducted to either identify or construct cellular pathways from a set of proteins. Pathways represent processes within cells or tissues, and by examining protein interactions, this analysis helps reveal diseases most statistically associated with the active protein chains.

Drug repurposing

Drug repurposing is an approach aimed at uncovering new therapeutic applications for approved or investigational drugs beyond their initial indications. Using the DiaBLE algorithm, i identified the top 20 putative disease genes and subsequently mapped these genes to associated drugs through the Drug-Gene Interaction Database (DGIdb). Then compiled a ranking based on the drugs most frequently associated with these 20 genes. This ranking was cross-referenced on ClinicalTrials.gov to determine whether these drugs are currently involved in clinical trials targeting obesity. The drugs with the highest associations were:

- **ACITRETIN**
- **TRETINOIN**
- **ALITRETINOIN**
- **ETRETINATE**

RESULTS AND DISCUSSION

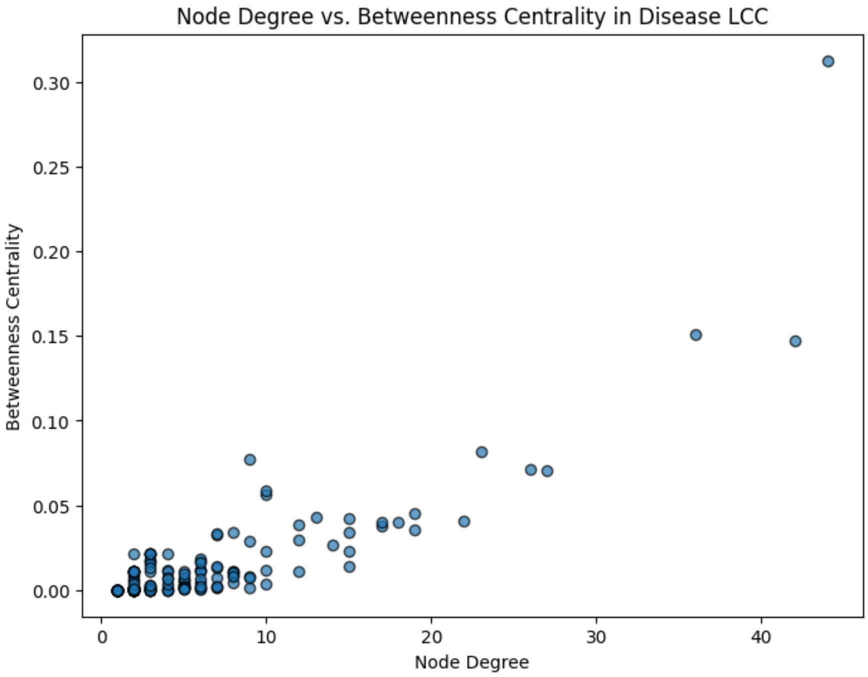
Table 1: Summary of GDAs and basic network data

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
Obesity	C0028754	C18	300	288	188

Table 2: Main network metrics of disease LCC genes

Ranking	GeneName	Degree	Betweenness	Eigenvector	Closeness	Ratio	
1	HSPA5	44	0.312774	0.276016	0.483204	0.007109	
2	EP300	42	0.147215	0.349896	0.46402	0.003505	
3	ESR1	36	0.151041	0.278897	0.452785	0.004196	
4	CEBPA	27	0.070459	0.234069	0.418345	0.00261	
5	PARP1	26	0.071181	0.228848	0.430876	0.002738	
6	NR3C1	23	0.081451	0.189671	0.424036	0.003541	
7	AR	22	0.040535	0.20924	0.412804	0.001843	
8	NCOA3	19	0.044936	0.17988	0.386364	0.002365	
9	SIRT1	19	0.035794	0.205138	0.40919	0.001884	
10	MTCH2	18	0.040116	0.109108	0.397872	0.002229	
11	AKAP1	17	0.038091	0.113786	0.403017	0.002241	
12	AKT1	17	0.04032	0.159139	0.406522	0.002372	
15	SMARCA4	15	0.02264	0.168921	0.385567	0.001509	
16	FASN	15	0.013946	0.128669	0.387967	0.00093	
14	CS	15	0.042575	0.038961	0.35619	0.002838	
13	EMC1	15	0.034221	0.139571	0.413717	0.002281	
17	PTPN1	14	0.026703	0.099068	0.370297	0.001907	
18	UQCRC2	13	0.043176	0.089542	0.393684	0.003321	
19	TFRC	12	0.038223	0.119481	0.417411	0.003185	
20	MYH9	12	0.029629	0.117201	0.404762	0.002469	
21	PPARG	12	0.011298	0.149723	0.397028	0.000941	
22	NPC1	10	0.056051	0.044753	0.349533	0.005605	
23	SOD1	10	0.022938	0.046299	0.349533	0.002294	
24	TF	10	0.058652	0.062735	0.371769	0.005865	
25	NR0B2	10	0.003193	0.114602	0.348881	0.000319	
26	NR1H2	10	0.011914	0.082008	0.322414	0.001191	
30	NR1H3	9	0.001253	0.104714	0.345018	0.000139	
31	FOS	9	0.007008	0.118232	0.363107	0.000779	
28	ADRB2	9	0.077051	0.044212	0.366667	0.008561	
29	HK1	9	0.028712	0.084681	0.367387	0.00319	
27	ACLY	9	0.007626	0.098449	0.37251	0.000847	
32	NUDC	8	0.008391	0.083667	0.363107	0.001049	

The scatter plot shows the relationship between Node Degree and Betweenness Centrality in the largest connected component of the disease network. Most nodes have low degrees and low betweenness, indicating limited connectivity and influence. A few nodes, however, have higher values, acting as potential hubs or key connectors. The outlier with both high degree and high centrality may play a critical role in disease-related pathways, highlighting its importance in the network structure**.**



Performance Comparison

The algorithms were evaluated through 5-fold cross-validation using precision, recall, and F1-score as metrics.

The figure below presents the error bars for each method across various numbers of selected genes. Performance comparisons indicate that DIAMOnD and DiaBLE perform similarly. However, according to findings in [3], the DiaBLE algorithm yields results with greater biological relevance compared to DIAMOnD. Therefore, despite comparable performance metrics, DiaBLE was selected as the optimal algorithm for identifying new putative disease genes.

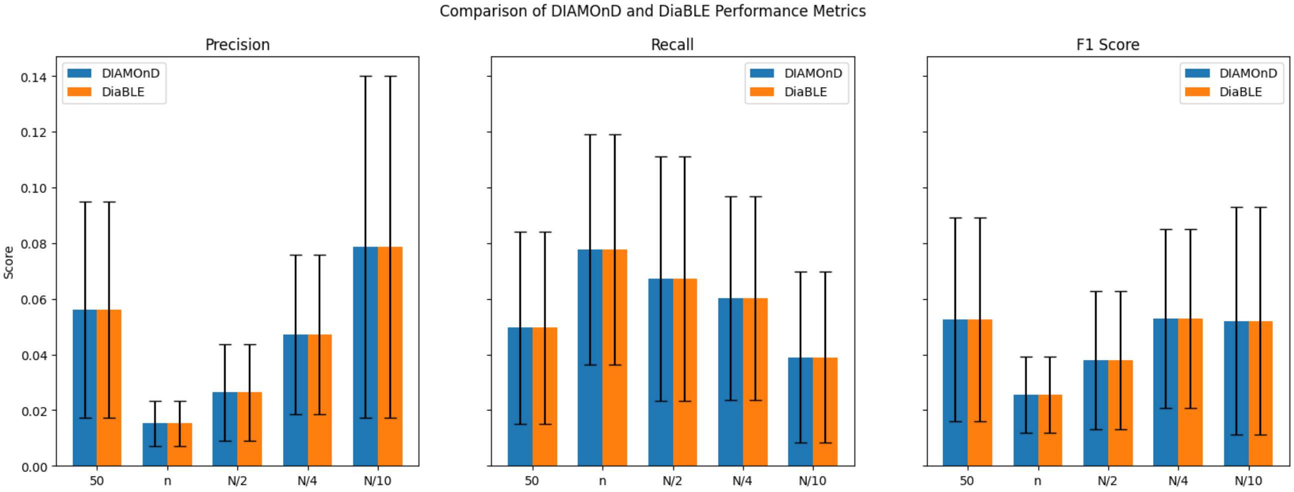


Figure 2: Performance Comparison between different algorithms for different numbers of selected genes

Analysis of Diffusion Algorithm Results

The diffusion-based algorithm produced ranked gene scores across different diffusion times (t=0.01, t=0.005, and t=0.002). These rankings highlight potential disease-associated genes, with scores reflecting the diffusion impact, which could indicate their importance or relevance within the disease gene network.

1. **Consistent Genes Across Times:**
   - The genes MIR130A, CCKAR, and GIPR consistently appear in the top three ranks across all diffusion times. This stability suggests that these genes are central within the diffusion network and might play critical roles in the disease's underlying biology.
2. **Changes in Scores and Rankings:**
   - The Score values slightly increase as the diffusion time decreases (t=0.002 shows the highest scores), likely due to tighter clustering around central nodes within shorter diffusion intervals. These incremental score differences indicate sensitivity to diffusion time, which could affect gene prioritization depending on parameter tuning.

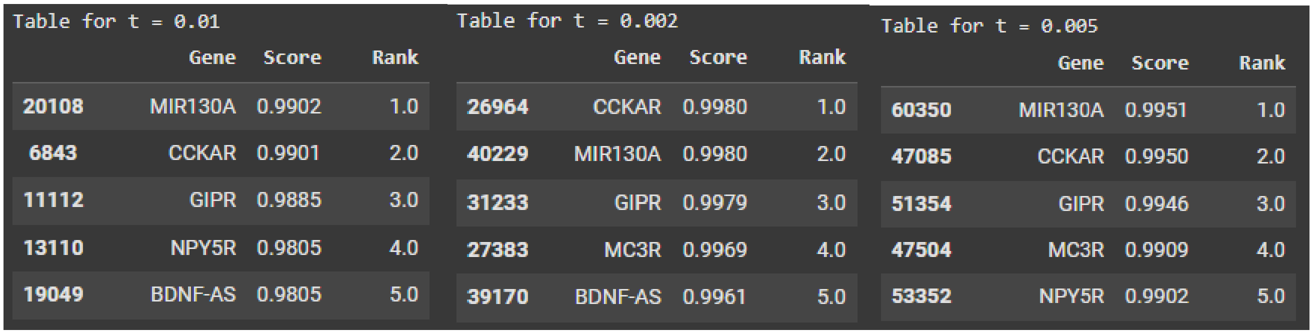


Table for t = 0.01

	Gene	Score	Rank
20108	MIR130A	0.9902	1.0
6843	CCKAR	0.9901	2.0
11112	GIPR	0.9885	3.0
13110	NPY5R	0.9805	4.0
19049	BDNF-AS	0.9805	5.0

Table for t = 0.002

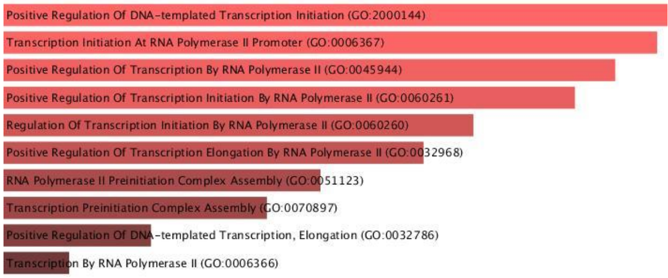
	Gene	Score	Rank
26964	CCKAR	0.9980	1.0
40229	MIR130A	0.9980	2.0
31233	GIPR	0.9979	3.0
27383	MC3R	0.9969	4.0
39170	BDNF-AS	0.9961	5.0

Table for t = 0.005

	Gene	Score	Rank
60350	MIR130A	0.9951	1.0
47085	CCKAR	0.9950	2.0
51354	GIPR	0.9946	3.0
47504	MC3R	0.9909	4.0
53352	NPY5R	0.9902	5.0

Enrichment Analysis

The figures below display the Enrichment Analysis results for the top 100 putative genes identified using DiaBLE. The bar plots indicate significance levels, with p-values representing the strength of association. The closer the p-value is to zero, the more significant the related GO term for the gene group, or, in the case of pathway analysis, the higher the probability of association with the disease.



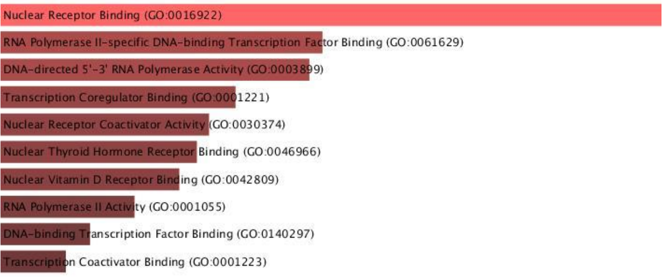
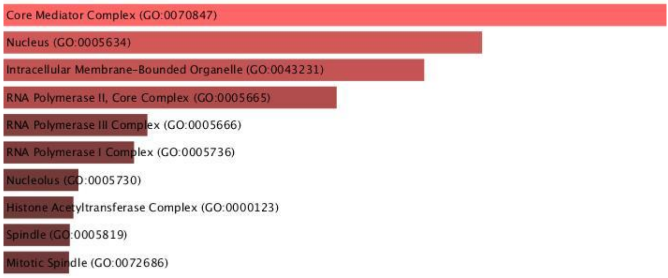


Figure 3: GO Biological Proces

Figure 4: GO Molecular Function



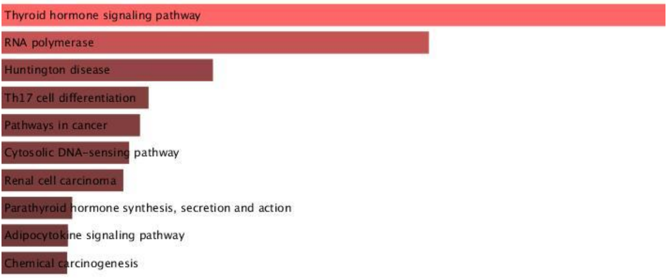
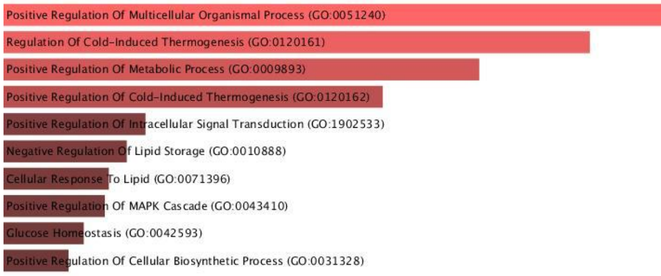


Figure 5: GO Cellular Component

Figure 6: KEGG Pathways

The enrichment analysis reveals that the top putative genes are significantly involved in transcriptional regulation, particularly within RNA polymerase II-associated processes (Figure 3) and nuclear receptor binding (Figure 4). Cellular component analysis (Figure 5) localizes these genes to the nucleus and transcriptional machinery. KEGG pathway analysis (Figure 6) links the genes to thyroid hormone signaling, cancer pathways, and other metabolic processes. These results suggest that the identified genes play crucial roles in gene regulation and may have broader implications in disease-related pathways.

Enrichment Analysis of All seed genes:



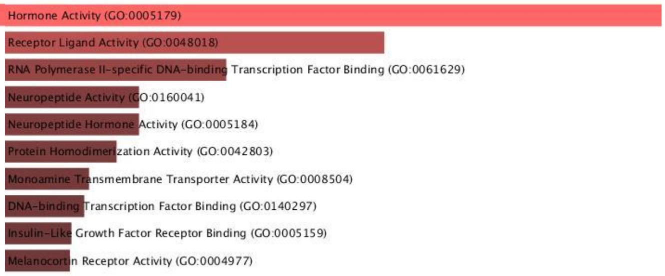
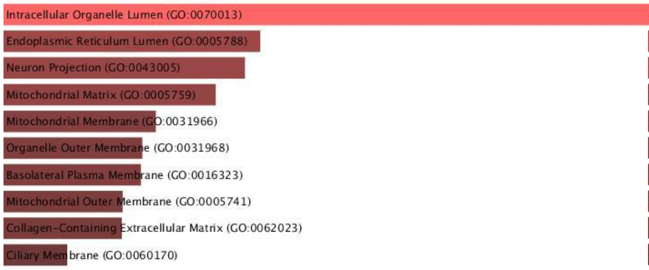


Figure 7: GO Biological Process

Figure 8: GO Molecular Function



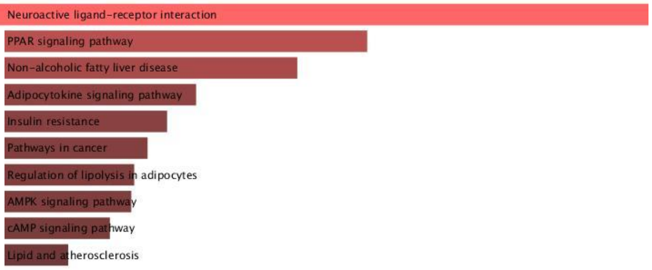


Figure 9: GO Cellular Component

Figure 10: KEGG Pathways

The enrichment analysis reveals that the seed genes are primarily involved in metabolic regulation, lipid storage, and glucose homeostasis (Figure 7), with significant roles in hormone activity and receptor binding for hormonal signaling and regulation (Figure 8). They are localized to intracellular organelles and mitochondria, indicating functions in metabolism and energy production (Figure 9). Key pathways include PPAR signaling, insulin resistance, and neuroactive interactions, linking these genes to metabolic and obesity-related disorders (Figure 10). These findings underscore the genes' roles in metabolic regulation, hormonal signaling, and energy-related processes associated with obesity.

REFERENCES

[1] Huang, T., Wang, L., Yan, Z., & Gao, Y. (2020). "Obesity-associated protein–protein interaction network: a pathway-driven network analysis of human obesity." *Journal of Cellular Biochemistry*, 121(2), 1344-1354. doi:10.1002/jcb.29364.

[2] Zhuang, Q., Shen, L., Ji, H., Zeng, H., & Wang, Y. (2021). "Protein-protein interaction network and pathway analysis to explore potential biomarkers for obesity-related diseases." Frontiers in Endocrinology, 12, 679345. doi:10.3389/fendo.2021.679345.

[3] M. Petti, D. Bizzarri, A. Verrienti, R. Falcone and L. Farina, "Connectivity Significance for Disease Gene Prioritization in an Expanding Universe," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 6, pp. 2155-2161, 1 Nov.-Dec. 2020, doi: 10.1109/TCBB.2019.2938512.