International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 6, December 2021, pp. 5240~5250
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i6.pp5240-5250

Hybrid approach for disease comorbidity and disease gene prediction using heterogeneous dataset

Lakshmi K. S.[1], Vadivu G.[2]
[1]Department of Computer Science Engineering, SRM Institute of Science and Technology (SRMIST), Kattankulathur, TamilNadu, India
[2]Department of Information Technology, SRM Institute of Science and Technology (SRMIST), Kattankulathur, TamilNadu, India

Article Info

Article history:

Received May 26, 2020
Revised May 18, 2021
Accepted Jun 12, 2021

Keywords:

Disease comorbidity
Gene ontology
Human phenotype ontology
Pathway
Protein-protein interaction
Random walk restart

ABSTRACT

High throughput analysis and large scale integration of biological data led to leading researches in the field of bioinformatics. Recent years witnessed the development of various methods for disease associated gene prediction and disease comorbidity predictions. Most of the existing techniques use network-based approaches and similarity-based approaches for these predictions. Even though network-based approaches have better performance, these methods rely on text data from OMIM records and PubMed abstracts. In this method, a novel algorithm (HDCDGP) is proposed for disease comorbidity prediction and disease associated gene prediction. Disease comorbidity network and disease gene network were constructed using data from gene ontology (GO), human phenotype ontology (HPO), protein-protein interaction (PPI) and pathway dataset. Modified random walk restart algorithm was applied on these networks for extracting novel disease-gene associations. Experimental results showed that the hybrid approach has better performance compared to existing systems with an overall accuracy around 85%.

Corresponding Author:

Lakshmi K. S.
Department of Computer Science Engineering
SRM Institute of Science and Technology (SRMIST)
Kattankulathur, TamilNadu, India
Email: lekshmy.shalu@gmail.com

1. INTRODUCTION

Disease comorbidity is defined as the existence of multiple disorders together with a primary disease [1]. This can add to the complexity of the treatment procedure and the condition of co-morbid patients is more complicated than that of patients suffering from any single disease. Comorbidity raises the difficulty of treating diseases that may potentially lead to higher mortality rates. Patients suffering from comorbid disease conditions need appropriate medical care and attention. Comorbidity can depend on pre-existing conditions or may exist as distinct conditions [2]. Though many diseases are comorbid, the underlying biological reasons remain obscured. Analyzing comorbid disease conditions by examining etiological components helps in discovering the basis mechanisms behind the development of diseases. Common genes, pathways and protein-protein interactions are biological factors contributing to disease comorbidities. The study of disease comorbidity can reveal the basic molecular disease mechanisms. This in turn can be useful for pinpointing disease causing genes and the associated biological pathways or vice-versa. Numerous research works have now been carried out on the basis of clinical data as well as molecular data for predicting disease comorbidities.

Most of these works rely on data mining techniques [3] for predicting novel comorbidities. Table 1 shows some of the existing systems available for disease comorbidity prediction.

Network approach based on random walks and statically approaches were mainly used for disease comorbidity prediction. A combination of statistical and network approach was also found to be effective in disease comorbidity prediction [4]-[6]. Another method, frequent pattern mining has also shown good performance for pattern analysis from outpatient records [7]. Post market adverse drug surveillance data was also used for comorbidity prediction [8].

Table 1. Existing systems for disease comorbidity predictions

Existing System	Techniques Used	Limitations
ComoR [9]	Statistical approach (Relative Risk and ϕ-correlation)	– Did not consider PPI and Pathway data
"Identification of disease comorbidity through hidden molecular mechanisms" [10]	Network-based approach (Random Walk Restart algorithm and XD Score)	– Protein-protein interaction data was not considered which also could contribute in detecting disease comorbidities. – Clinical data-based test was not conducted
Comorbidity [11]	Network-based approach (Random Walk Restart algorithm)	– Did not consider PPI and Pathway data
"Finding disease similarity based on implicit semantic similarity" [12]	Semantic similarity using GO terms	– Causal factors not considered – Impossible to discriminate the primary disease and the comorbid disease
PCID [1]	Similarity based approach	– The pathway information considered was limited to the current knowledge on molecular pathways, and more signals may be discovered if the PPI network was used instead of current pathway annotations. – When the multi-view similarity was integrated based on different types of data, contribution from each data was assumed to be equal. – Disease comorbidities exist in a condition-specific manner. Data contributing to this fact like age and sex was not considered during the development of PCID.
CORE [13]	Clustering and Association Analysis	– Considered only past medical history of patients – Did not consider current symptoms and clinical or biological factors – Dataset was small

Most of the existing systems, consider only one or two biological factors underlying the comorbidity patterns such as genes, pathways, biological process, cellular component, protein-protein interaction and molecular functions for disease comorbidity prediction. Integration of these data helps in increasing the predictive power of prediction methods. Latest researches in disease comorbidity prediction was successful in integrating multi-scale dataset. But all the datasets were given equal weightage while making comorbidity predictions.

In this work, a dataset ranking algorithm has been used for finding the relevance of each dataset in comorbidity prediction. Using weighted association rule mining, disease comorbidity patterns were generated. CTD [14] database contains information on disease-gene associations, disease-pathway associations and phenotype-disease interactions. These information were mainly used for disease comorbidity prediction. Additional databases used in this work is discussed in the next section.

2. DATASET USED

Three types of data have been considered: PPI, Pathway and Gene Ontology annotations in the proposed method. PPI data were downloaded from MINT [15], HPRD [16] and IntAct [17]. CTD and DisGeNet [18] databases were used for obtaining disease-gene associations. A total of 39239 PPI were taken for conducting experiment and a maximum of up to 234 diseases were found to be associated with each PPI. The two main sources of dataset for disease-pathway information has been: "Molecular Signatures DataBase" (MSigDB) 4.0 and CTD. A set of 1077 pathways were obtained from "curated(c2) gene sets" in MSigDB V4.0.About 551548 pathways were taken from CTD database. For the task of association rule mining, a transaction corresponds to a pathway in the pathway database and the diseases related to the pathway were considered as data items belonging to the respective transaction. A total of 2332 pathways were found to have "comorbid disease conditions" associated with them. CTD database contains "diseases-GO annotations" mappings that can be directly extracted for mining task. There are 118773 associations between diseases and GO cellular components, 145579 associations between Diseases and GO molecular function and 671095 associations between diseases and GO biological process.

3. PROPOSED METHOD

The overall architecture of the system is shown in Figure 1. In the existing available system for disease comorbidity prediction using diverse dataset, similarity based approach was used. The disadvantage of this technique is that all the datasets were given equal weightage or their contribution in comorbidity prediction was assumed to be equal. To overcome this disadvantage, we introduced a dataset ranking algorithm [19] which is based on multi-criteria decision analysis. Three different data set were considered: protein-protein interaction (PPI) data, Pathway data and gene ontology (GO) [20], [21] annotation data. Since these datasets are diverse in nature, their contribution in comorbidity prediction will be different. So there exists the need to evaluate the dataset based on their contribution in comorbidity prediction.

Multi-criteria decision analysis approach was chosen for ranking datasets. Multi-criteria decision analysis is a ranking method that is commonly used to arrange a finite number of decision alternatives, each of which is clearly described in terms of different characteristics. These characteristics are also often called attributes or decision criteria. Selection of criteria for decision analysis is purely based on user requirement. Accuracy of a system is defined in terms of true positive, true negative, false positive and false negative values. Since we are focusing on accuracy of prediction, the decision criteria considered for dataset ranking are true positive, true negative, false positive and false negative values. Out of these, true positive and true negative values are more relevant since they directly contribute to the expected results for prediction whereas false values have no direct contribution to prediction results. They are only used for finding accuracy. These values are calculated using existing approach [1] based on similarity measurement. Pathway dataset is ranked highest as per the decision analysis. Pseudocode for dataset ranking algorithm is given below:

Input:
PPI Data-D_1,
Pathway Data-D_2,
Gene Ontology Annotations-D_3,
Primary Disease PD, Query Disease QD,
Performance Parameters-TP, TN, FN
Associated weights W_1, W_2, W_3 as 3, 2, 1 for TP,
TN and FN respectively
Steps:
Calculate disease similarity between PD and QD using D_1, D_2, D_3

```
for i in 1 to 3
find TP(Di), TN(Di), FN(Di)
for i in 1 to 3
```

$$C(Di,Dj) = \frac{\sum \forall j: g(Di) \geq g(Dj) Wj}{\sum_{j=1}^{n} W\,j}$$

$$If\ \forall x\ gx(Di) \geq gx(Dj) then$$

$$D(Di,Dj) = 0$$

```
else
```

$$D(Di,Dj) = \frac{1}{\delta} max[gy(Dj) - gy(Di)]$$

$$If\ C(Di,Dj) \geq Ct\ and\ D(Di,Dj)) \leq dt\ then$$

```
Di outperforms Dj;
```

Rank Di accordingly with the highest weight Wx.

After ranking the dataset, weighted association rule mining was performed for generating disease comorbidities. In the next phase, disease comorbidities were generated using network-based method. Network based approaches are usually used for the prediction of disease associated genes. Variants of random walk restart (RWR) algorithm are applied for finding disease associated genes from disease-gene networks. RWR starts randomly from a seed node to another, but can also restart the navigation in a new arbitrary node. Thereby, depending on the topological structure of the network, some nodes will be visited more frequently than others. The number of visits is considered as a proxy measure of relevance of each node with respect to seed node. In our work, instead of directly using disease network and gene network, HPO [22] and GO annotations were used for constructing network. GO annotations and HPO annotations were downloaded from gene ontology and human phenotype ontology. Using weighted association rule mining, GO term associations and HPO term associations were obtained [23]-[26]. Two networks were constructed; one containing nodes of GO terms and other having nodes of HPO terms. These two networks were then connected using known GO-HPO associations. Thus a bipartite graph was formed. Then a modified random walk was performed on this bipartite graph to generate novel disease-gene correlation and disease comorbidities [27].

Rules generated in the first two phases were combined to generate comorbidity network with nodes, diseases and genes. A query disease and target disease were given as input to the system. Then a query-

disease vector and target-disease vector were generated with known disease-gene association. RWR was repeatedly performed on the hybrid network with query disease as seed node and restart probability of 0.8. Then the cosine similarity was calculated between query disease vector and target disease vector.

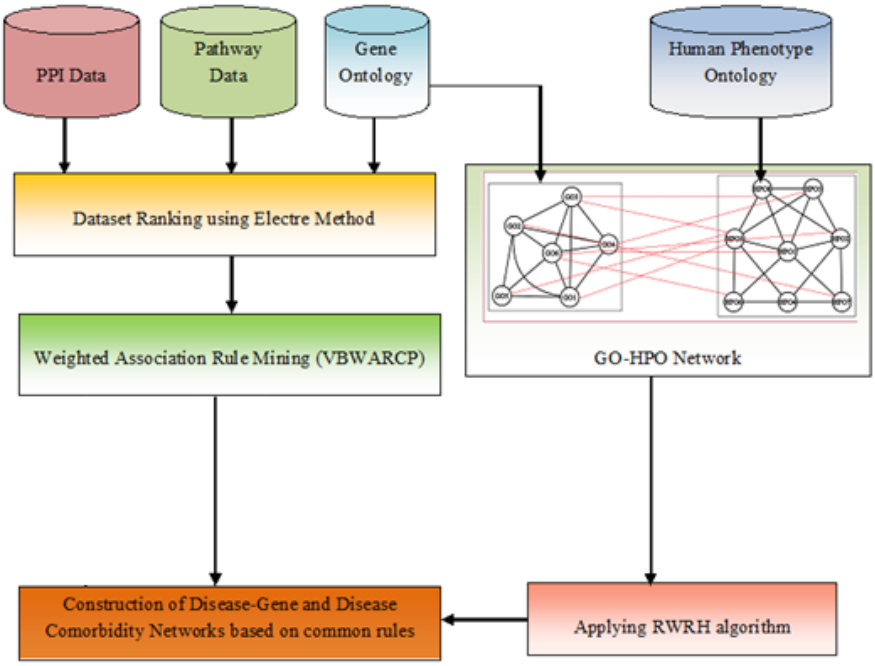

Figure 1. Overall system architecture

Algorithm for the proposed method is given below:

HDCDGP Algorithm
Input: PPI Data, Pathway Data, GO Annotations, HPO Annotations, Query Disease
Output: Novel disease comorbidities and gene-disease associations
Steps:

1. Start
2. Collecting PPI data, Pathway data and Gene Ontology data from PPI and CTD databases
3. Ranking the dataset based on multi-criteria decision analysis using ELECTRE – I method
4. Calculate purity of data items as $P_k = 1 - \frac{log_2(|N_k|)}{log_2(|I|)^2}$
5. Calculate connectivity of data items as $l_k = \sum_i^n \frac{count(ik)}{count(k)}$
6. Calculate valence weight as

$$V_k = \delta.\ P_k + (1-\delta) \sum_i^n \frac{count(ik)}{count(k)} .\ P_i$$

7. Define weighted support as the product of dataset rank and valence weight
8. Weighted Association rule mining is performed to generate novel disease comorbidities
9. Perform weighted association rule mining in Gene Ontology to generate GO term associations.
10. Perform weighted rule mining in HPO to generate HPO term associations.
11. Generate GO network and HPO network based on the association rules.
12. Combine the networks using known GO-HPO associations to generate a bi-partite graph.
13. Apply random walk restart algorithm to obtain novel disease-gene correlation.
14. Construct a hybrid network with edges that are associations of the form disease-disease, disease-gene and gene-gene.
15. A disease vector is generated for query disease with known disease-gene interactions.
16. Perform random walk on the hybrid network with a restart probability of 0.8.
17. Generate disease vectors for target diseases.

18. Compute cosine similarity as

$$Sim(D_q, D_t) = \frac{D_q . D_t}{\|D_q\| \|D_t\|} = \frac{\sum_{i=1}^{n} D_{qi} D_{ti}}{\sqrt{\sum_{i=1}^{n} {D_{qi}}^2} \sqrt{\sum_{i=1}^{n} {D_{ti}}^2}}$$

19. Return target disease when cosine score is above the given threshold.
20. Stop.

4. RESULTS AND DISCUSSION

The network statics of the hybrid network is given in Table 2. Figures 2 to 7 demonstrates the comparison of various network parameters of the resultant network (graph) generated using Cytoscape [28]. From Figure 2, it can be seen that the maximum number of shared neighbors is less than 20. The shortest path length distribution in Figure 3 gives the number of node pairs (n, m) with various shortest path length. The graph shows that shortest path length of 4 is maximum for a frequency of 400000. Closeness centrality is another network parameter that describes how fast a node can be approached (reachable) from other nodes. It is the inverse of average shortest path length. Figure 4 depicts closeness centrality versus number of neighbors. Node degree distribution gives the number of edges connected to a given node. Figure 5 shows the node degree distribution of the resultant graph. If a node is connected to more number of nodes, the comorbidity information related to that node may be less accurate; since the number of such nodes are less as evident from the graph, the comorbidity information is more reliable. Figure 6 illustrates the neighborhood connectivity distribution of the graph. The neighborhood connectivity of a node is the average connectivity of all neighbors of that node and the neighborhood connectivity distribution gives the average of the neighborhood connectivities of all nodes in the given graph.

Table 2. Network statistics

Parameter	Value
Clustering Coefficient	0.154
Connected Components	81
Network Centralization	0.113
Average number of neighbors	11.756
Number of Nodes	10014
Network Density	0.001
Network Heterogeneity	1.878
Number of self-loops	1
Multi-edge node pairs	2
Analysis time (sec)	0.919

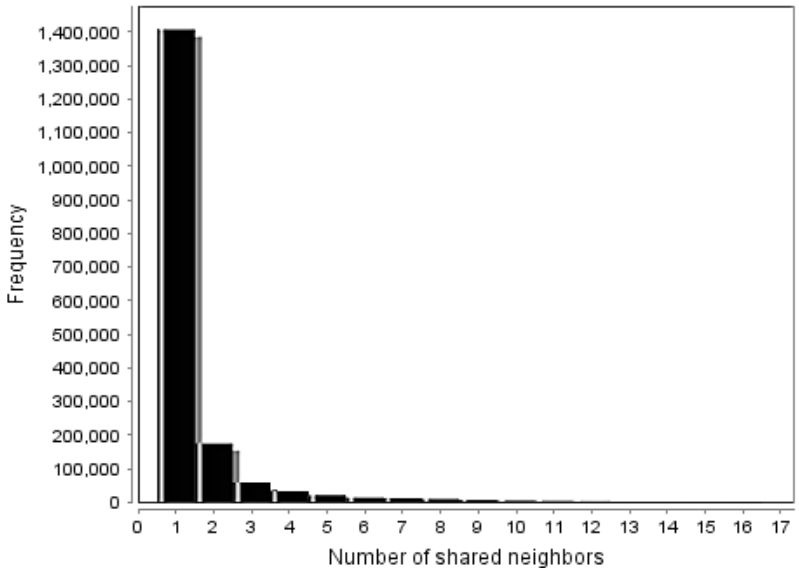

Figure 2. Shared neighbors' distribution

Clustering coefficient of a node n is defined as:

Cn=2en/(kn (kn-1)), where kn is the number of neighbors of n and en is the number of connected pairs between all neighbors of n. Figure 7 illustrates the average clustering coefficient distribution of the graph. 81 connected components are present in the resultant graph.

Case Study: MESH: D000544 (Alzheimer's disease)

Comorbid conditions associated with Alzheimer's disease were studied in detail. There were about 663 pathways found to be associated with Alzheimer's disease. 1824 biological processes, 324 cellular components and 493 molecular functions in gene ontology were found to be associated with Alzheimer's disease. Tables 3, 4, 5 and 6 show the comorbidity supporting statistics of Alzheimer's disease with Schizophrenia, coronary artery disease, dementia, bipolar disorder and diabetes mellitus.

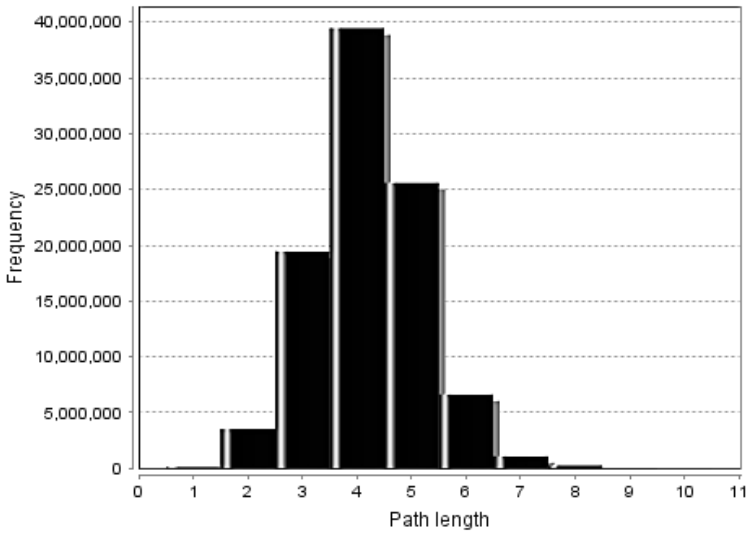

Figure 3. Shortest path length distribution

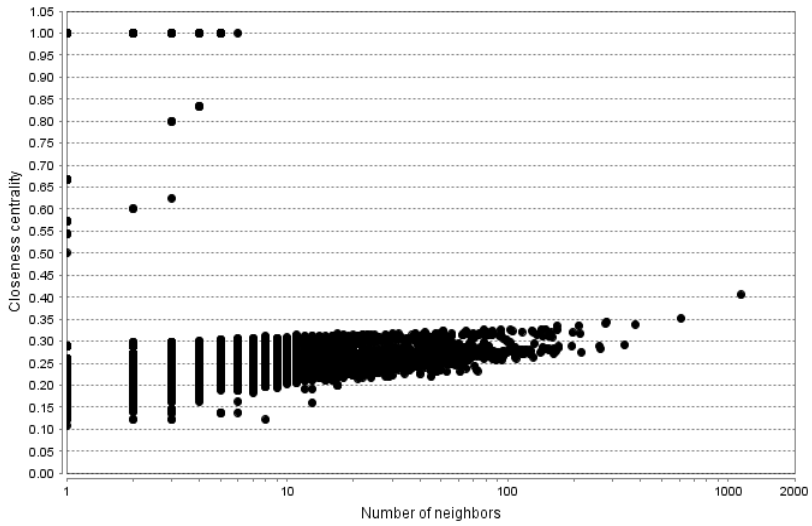

Figure 4. Closeness centrality

For evaluating the performance of proposed method, a gold standard set of 14 diseases were chosen: HP:0100615 (Ovarian Cancer), HP:0001658 (Myocardial Infarction), HP:0001250 (Epilepsy), HP:0002511 (Alzheimer's Disease), HP:0000822 (Hypertension), HP:0002099 (Asthma), HP:0002665 (Lymphoma), HP:0001638 (Cardiomyopathy), HP:0000821 (Hypothyroidism), HP:0001875 (Neutropenia), HP:0000726 (Dementia), HP:0001370 (Rheumatoid arthritis), HP:0001909 (Leukemia), HP:0001677 (Coronary artery disease).

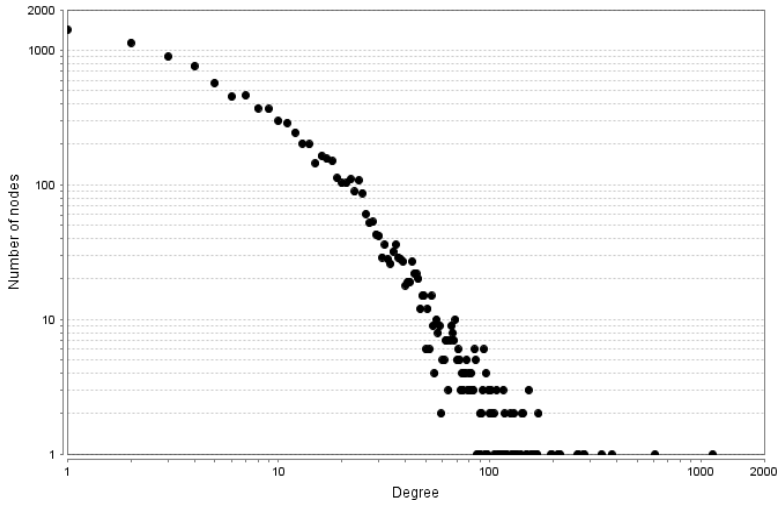

Figure 5. Node degree distribution

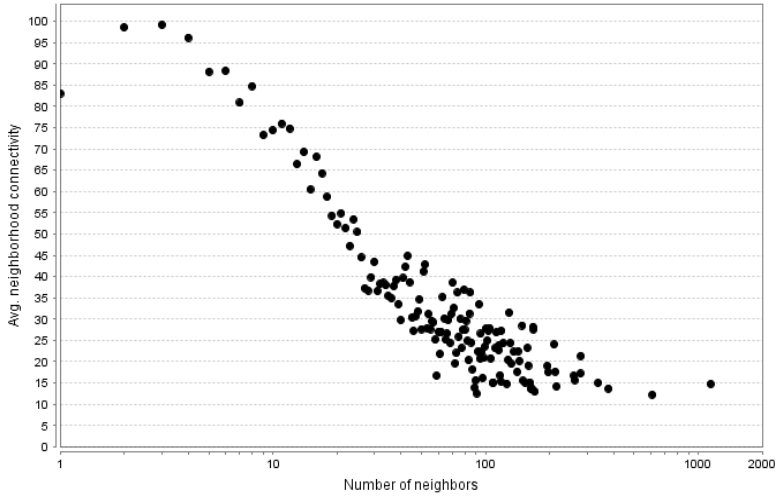

Figure 6. Neighborhood connectivity distribution

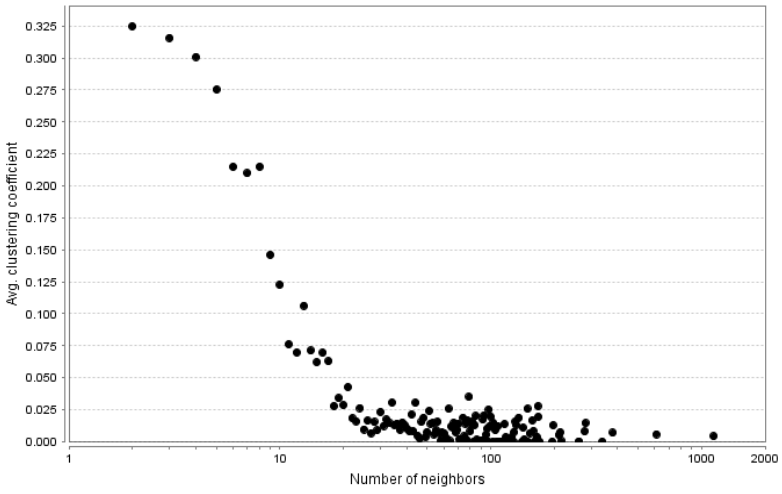

Figure 7. Average clustering coefficient distribution

Table 3. Alzheimer's disease-comorbidity statistics based on cellular components

Comorbid Disease	Number of common cellular components (GO Terms)	Examples of common cellular components (GO Terms)
Schizophrenia	103	GO:0001669
		GO:0005912
		GO:0031362
		GO:0031225
		GO:0046658
Coronary Artery Disease	100	GO:0005884
		GO:0032432
		GO:0005912
		GO:0097208
		GO:0045177
Dementia	80	GO:0032432
		GO:0097208
		GO:0106003
		GO:0046658
		GO:0045179
Bipolar Disorder	132	GO:0015629
		GO:0097208
		GO:0031225
		GO:0046658
		GO:0097440
Diabetes Mellitus	92	GO:0005884
		GO:0032432
		GO:0005912
		GO:0097208
		GO:0046658

Table 4. Alzheimer's disease-comorbidity statistics based on molecular functions

Comorbid Disease	Number of common molecular functions (GO Terms)	Examples of common molecular functions (GO Terms)
Schizophrenia	117	GO:0004115
		GO:0033130
		GO:0004559
		GO:0034185
		GO:0034618
Coronary Artery Disease	207	GO:0003990
		GO:0003993
		GO:0003779
		GO:0003785
		GO:0022853
Dementia	126	GO:0004115
		GO:0033130
		GO:0003993
		GO:0003994
		GO:0003779
Bipolar Disorder	187	GO:0004115
		GO:0003993
		GO:0003994
		GO:0003779
		GO:0051015
Diabetes Mellitus	171	GO:0004115
		GO:0003990
		GO:0033613
		GO:0004017
		GO:0004935

To validate the performance Leave one out cross validation (LOOCV) is applied. The ROC curve is given in Figure 8. To avoid mismatch in disease id as different datasets follow different ID for the same disease, MESH ID was chosen in common. The area under curve (AUC) for the ROC curve given in Figure 8 was calculated which turned up to 0.853. Figure 9 illustrates the performance comparison of proposed method with existing approaches: PCID [1], VBWARCP [19], OBDCP [27].

Table 5. Alzheimer's disease-comorbidity statistics based on biological process in gene ontology

Comorbid Disease	Number of common biological process (GO Terms)	Examples of common Biological Process (GO Terms)	PMID
Schizophrenia	306	GO:0090630	26312426
		GO:0000185	
		GO:0070162	
		GO:0046032	
		GO:0097113	
Coronary Artery Disease	965	GO:0019471	16918818
		GO:0019857	
		GO:0008292	
		GO:0006581	
		GO:0001507	
Dementia	516	GO:0008292	
		GO:0006581	26312426
		GO:0006085	
		GO:0090527	
		GO:0019857	
Bipolar Disorder	758	GO:0095500	28476640
		GO:0007015	
		GO:0001508	
		GO:0007190	
		GO:0097202	
Diabetes Mellitus	917	GO:0019471	30542257
		GO:0019857	
		GO:0008292	
		GO:0006581	
		GO:0095500	

Table 6. Alzheimer's disease-comorbidity statistics based on pathways

Comorbid Disease	Number of common pathways	Examples of common pathways
Schizophrenia	233	REACT:R-HSA-111447
		REACT:R-HSA-114452
		REACT:R-HSA-451308
		REACT:R-HSA-211227
		REACT:R-HSA-451326
Coronary Artery Disease	325	REACT:R-HSA-382556
		KEGG:hsa02010
		REACT:R-HSA-1369062
		REACT:R-HSA-166054
		REACT:R-HSA-451326
Dementia	82	KEGG:hsa05034
		KEGG:hsa05010
		KEGG:hsa05031
		REACT:R-HSA-977225
		REACT:R-HSA-109581
Bipolar Disorder	282	REACT:R-HSA-166054
		REACT:R-HSA-451308
		REACT:R-HSA-451326
		REACT:R-HSA-1280218
		KEGG:hsa04920
Diabetes Mellitus	241	REACT:R-HSA-1280218
		KEGG:hsa04520
		KEGG:hsa04920
		KEGG:hsa04261
		KEGG:hsa04933

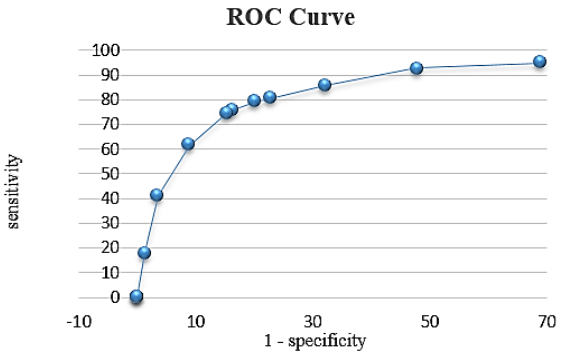

Figure 8. ROC curve

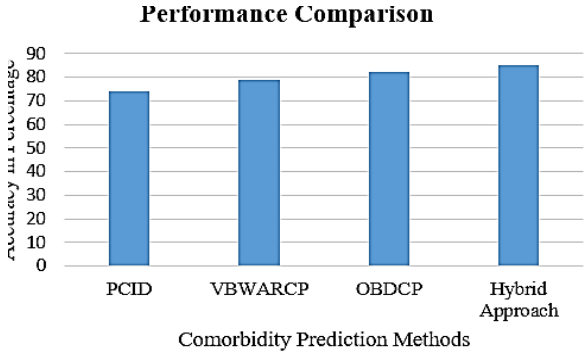

Figure 9. Performance comparison

In this work, we have developed a weighted association rule mining method for discovering disease comorbidities from heterogeneous forms of dataset. By implementing a dataset ranking algorithm, the performance of algorithm on various datasets like protein-protein interaction (PPI) data, pathway data and gene ontology dataset were studied. Rather than providing the datasets the same weightage, PPI dataset is given the highest rank, based on multi-criteria decision analysis. This could help in increasing the accuracy of predicting disease comorbidities. Valence based weighted association rule mining for disease comorbidity prediction (VBWARCP) algorithm has been developed based on weighted Apriori algorithm by introducing weights in terms of dataset rank, purity and linkage. VBWARCP has got an accuracy of 79%.

In the second phase, a bipartite network is developed with terms from gene ontology (GO) and human phenotype ontology (HPO). Gene ontology terms were connected using associations obtained from gene ontology by applying weighted association rule mining algorithm. Similarly, HPO terms were connected using associations among HPO terms derived as association rules. These rules were generated from human phenotype ontology by means of weighted association rule mining. Then the terms in Gene Ontology were mapped to terms in human phenotype ontology terms using known associations. These relationships were extracted from human phenotype ontology (HPO). HPO website contains HPO-gene associations. Thus an ontology based system for disease-comorbidity prediction (OBDCP) was developed. Application of random walk restart on heterogeneous network (RWRH) on the resultant network helps in finding disease comorbidities as well as disease associated genes. Accuracy of OBDCP was around 77%. Though it was less than the accuracy of VBWARCP, it was higher than the accuracy of similarity based approach.

Using the patterns and associations obtained in the first two phases, a comorbidity network has been constructed. The network consists of disease-disease edges (using known disease-disease associations and the result obtained using VBWARCP and OBDCP), gene-gene edges (constructed using PPI data) and disease-gene edges (constructed using already known gene-disease associations and those obtained using OBDCP). If a medical record is given as input to this system, medical terms can be extracted and the resultant diseases can be given as query diseases and the comorbid diseases associated with a patient can be found. It is also possible to find the genes associated with the query disease.

5. CONCLUSION

HDCDGP presents a reliable method to study disease comorbidities that can be suggested for high-throughput and clinical data analysis. Causal inference of diseases can be learned by the analysis of disease comorbidities and disease gene associations. Compared to the existing systems, our approach has gained an overall accuracy of 85%. HDCDGP is capable of finding novel disease comorbidities as well as disease-gene correlation. This approach will guide the researchers in improved understanding of the complex pathogenesis of disease risk phenotypes and the heterogeneity of diseases.

Discovering associations among diseases provides a deep cognizance on the underlying causes of diseases and assists in the task of disease-associated gene prediction and also in the development of novel drugs. Since associations among genes are not the only factor contributing to disease comorbidities, information of common pathways and abnormalities in cellular components, biological processes and metabolic functions in genes in detecting comorbid diseases will increase the prediction accuracy. If doctors are aware of the possibility of various comorbid disease conditions, it could help them in choosing treatment methods and also alleviates the sufferings that the patients may undergo. The proposed model can be further expanded by considering miRNA expressions. Adding more molecular level information will help in increasing the accuracy as well as discovering novel comorbidities. The system can also be used for developing new drugs and treatment procedures for comorbid disease conditions.

REFERENCES

[1] F. He, G. Zhu, W. Yin-Ying, Z. Xing-Ming and H. De-Shuang, "PCID: A Novel Approach for Predicting Disease Comorbidity by Integrating Multi-Scale Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 14, no. 3, pp. 678-686, 2017, doi: 10.1109/TCBB.2016.2550443.

[2] R. Jones, "Chronic Disease and Comorbidity," *British Journal of General Practice,* vol. 60, no. 575, 2010, Art. no. 394, doi: 10.3399/bjgp10X502056.

[3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Second Edition, Morgan Kaufmann, 2006.

[4] C. Di *et al.,* "Recognition of Disease Comorbidity Medication Patterns Based on Network Motif Analysis," *Research and Reviews: Journal of Pharmacy and Pharmaceutical Sciences,* vol. 5, no. 3, 2016.

[5] A. Khan, S. Uddin and U. Srinivasan, "Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression," *International Journal of Medical Informatics,* vol. 115, pp. 1-9, doi: 10.1016/j.ijmedinf.2018.04.001.

[6] S. Adnan and D. Alalwan, "Diabetic analytics: proposed conceptual data mining approaches in type 2 diabetes dataset," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS),* vol. 14, no. 1, pp. 88-95, 2019, doi: 10.11591/ijeecs.v14.i1.pp88-95

[7] S. Boytcheva, G. Angelova, Z. Angelov and D. Tcharaktchiev, "Mining comorbidity patterns using retrospective analysis of big collection of outpatient records," *Health Information Science and Systems,* vol. 5, no. 1, 2017, Art. no. 3, doi: 10.1007/s1375 5-017-0024-y.

[8] C. Zheng and R. Xu, "Large-scale mining disease comorbidity relationships from post-market drug adverse events surveillance data," *BMC Bioinformatics,* vol. 19, no. Suppl 17, 2018, Art. no. 500, doi: 10.1186/s12859-018-2468-8.

[9] M. A. Moni and P. Liò, "comoR: a software for disease comorbidity risk assessment," *Journal of Clinical Bioinformatics,* vol. 4, no. 8, 2014, doi: 10.1186/2043-9113-4-8.

[10] Y. Ko, M. Cho, J.-S. Lee and J. Kim, "Identification of disease comorbidity through hidden molecular mechanisms," *Scientific Reports,* vol. 6, 2016, Art. no. 39433, doi: 10.1038/srep3 9433.

[11] A. G.-Sacristan, A. Bravo, A. Giannoula, M. A. Mayer, F. Sanz and L. I. Furlong, "comoRbidity: an R package for the systematic analysis of disease comorbidities," *Bioinformatics,* vol. 34, no. 18, pp. 3228-3230, 2018, doi: 10.1093/bioinformatics/bty315.

[12] S. Mathur and D. Dinakarpandian, "Finding disease similarity based on implicit semantic similarity," *Journal of Biomedical Informatics,* vol. 45, no. 2, pp. 363-371, 2012, doi: 10.1016/j.jbi.2011.11.017.

[13] F. Folino and C. Pizzuti, "A Comorbidity-based Recommendation Engine for Disease Prediction," *2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS),* 2010, doi:10.1109/cbms.2010.6042664.

[14] P. A. Davis *et al.,* "The Comparative Toxicogenomics Database: update 2019," *Nucleic Acids Research,* vol. 47, no. D1, pp. D948-954, 2018, doi: 10.1093/nar/gky868.

[15] L. Licata *et al.*, "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Research*, vol. 40, pp. D857-D61. 2011, doi: 10.1093/nar/gkr930.
[16] K. T. S. Prasad *et al.*, "Human Protein Reference Database-2009 Update," *Nucleic Acids Research*, vol. 37, pp. D767-D72, 2009, doi: 10.1093/nar/gkn892.
[17] S. Orchard *et al*., "The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Research*, vol. 42, no. D1, pp. D358-D363, 2014, doi: 10.1093/nar/gkt1115.
[18] J. Piñero *et al.*, "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, vol. 48, no. D1, pp. D845-D855, 2019, doi:10.1093/nar/gkz1021.
[19] S. K. Lakshmi, G. Vadivu, "A Novel Approach for Disease Comorbidity Prediction Using Weighted Association Rule Mining," *Journal of Ambient Intelligence and Humanized Computing*, 2019, doi: 10.1007/s12652-019-01217-1.
[20] M. Ashburner *et al.*, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25-29, 2000.
[21] The Gene Ontology Consortium, "The Gene Ontology Resource: 20 years and still Going strong," Nucleic Acids Research, vol. 47, no. D1, pp. D330-D338, 2019, doi: 10.1093/nar/gky1055.
[22] S. Köhler *et al.*, "Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1018-D1027, 2018, doi: 10.1093/nar/gky1105.
[23] G. Agapito, M. Cannataro, P. H. Guzzi, and M. Milano, "GO-WAR: A Tool for Mining Weighted Association Rules from Gene Ontology Annotations," *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics-CIBB 2014*, vol. 8623, 2015, pp. 3-18, doi: 10.1007/978-3-319-24462-4 1.
[24] P. H. Guzzi, G. Agapito, M. Milano, and M. Cannataro, "Learning Weighted Association Rules in Human Phenotype Ontology," *arXiv:1701.00077*, 2016.
[25] S. Tongphu, B. Suntisrivaraporn, and P. Aimmanee, "Toward Semantic Similarity Measure between Concepts in an Ontology," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 14, no. 3, pp. 1356-1372, 2019, doi: 10.11591/ijeecs.v14.i3.pp1356-1372.
[26] R. Mohemad, F. Akma, N. Maizura, and A. A. Che, "The development of an ontology model for early identification of children with specific learning disabilities," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 5486-5494, 2019 doi: 10.11591/ijece.v9i6.pp5486-5494.
[27] S. K. Lakshmi, and G. Vadivu, "Network based approach for discovering novel gene-phenotypic association and disease comorbidities using ontological data," *Procedia Computer Science*, vol. 167, pp. 819-829, 2020, doi: 10.1016/j.procs.2020.03.421.
[28] H. Xu, M. A. Moni, and P. Lio, "CytoCom: A Cytoscape app to visualize, query and analyse disease comorbidity networks," *Bioinformatics*, vol. 31, no. 6, pp. 969-971, 2015, doi: 10.1093/bioinformatics/btu731.

BIOGRAPHIES OF AUTHORS

Lakshmi K. S. was born on 15th March 1984 at Cherai near Kochi, Kerala, India. She did her schooling in Lobelia English Medium High School. She obtained her B.Tech degree in Computer Science and Engineering from College of Engineering, Kidangoor. She pursued M.Tech in Computer and Information Science from Cochin University of Science and Technology, Kerala. She is currently a Part-time Research Scholar in the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. Her research interest includes Data Mining and Bioinformatics. Presently, she is also working as Assistant Professor in the Department of Information Technology, Rajagiri School of Engineering & Technology, Kochi affiliated to A.P.J Abdul Kalam Technological University.

G. Vadivu was born on 4th April 1972 at Sethiathope near Chidambaram. She did her schooling in DGM higher secondary school, Sethiathope. She obtainedher B.E. Computer Science and Engineering from Institute of Road and TransportTechnology, Bharathiar University. She did her M.Tech in Computer Science and Engineering at SRM University. After her under graduate degree she worked as a Lecturer for seven years. Then she joined SRM University during May 2000, andnow she is working as Professor and Head of the Department of Information Technology. Under Ph.D program, she carried out research on title "Semantic Similarity Measures to find Mapping and Relatedness of Terms Using Ontology". She has more than 30 publications in reputed International Journals and has participated and presented papers in several National and International conferences. She won the best teaching faculty award at SRM Institute of Science and Technology.