VECTOR BASED APPROACH FOR DISEASE COMORBIDITY PREDICTION USING HETEROGENEOUS LARGE SCALE DATASET

Lakshmi K.S*

Department of Computer Science Engineering,
SRM Institute of Science and Technology,
Kattankulathur, TamilNadu, India
E-mail: lekshmy.shalu@gmail.com†

Vadivu G

Department of Information Technology,
SRM Institute of Science and Technology,
Kattankulathur, TamilNadu, India
E-mail: vadivug@srmist.edu.in

Abstract

Generation of large scale biological data paved the development of novel methods for the discovery of underlying reasons behind disease development and progression. Disease comorbidity prediction and disease associated gene prediction has gained much importance during the last few decades. Interaction among genes, pathways, biological processes, molecular functions and cellular components are considered as the prominent biological factors which have causative roles in disease development. Integration of these heterogeneous data will help in finding disease comorbidities and disease gene prediction with improved accuracy. In this paper, a vector based approach has been proposed for finding novel comorbidities by integrating heterogeneous dataset such as PPI (protein-protein interaction), Pathway and Gene Ontology information including biological process, molecular function and cellular components. This study demonstrated that the integration of direct as well as indirect interaction among genes and high-level molecular association can be exploited for discovering significantly strong conditions of comorbidity.

***Keywords*:** Disease comorbidity; protein-protein interaction; pathway; Gene Ontology; Vector based similarity.

1. Introduction

Diseases have always perplexed mankind since time immemorial; andscientists the world over, have been grappling with solutions to all themaladies that plague us. This essentially introduces us to the topic of‘Comorbidity Research”. Comorbidity refers to the existence of one or more diseases along 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. An elucidation of pathological properties of varied diseases and theircoordinated activities at the molecular level is what ComorbidityResearch is all about. The 21st century has increased our awareness ofhuman disease mechanisms thus providing ample evidence thatcomplex diseases stem from the breakdown of concerted activities ofgenes involved in common or related cellular processes. 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, statistical approaches and similarity-based approaches for these predictions. PCID[2], ComoR[3] and Comorbidity[4] are some of the existing systems available for comorbidity prediction. ComoR uses statistical method for finding disease comorbidity whereas Comorbidity is a network based approach. PCID, the latest technique make use of similarity based approach for finding novel comorbidities.

*Assistant Professor, Department of Information Technology, Rajagiri School of Engineering and Technology, Kochi, Kerala, India
† lakshmiks@rajagiritech.edu.in.

Thecoexistence of two or more diseases in an individual raises the questionabout their underlying common etiological pathways. Hence thecomorbidity patterns of diseases can help us comprehend the basicmolecular disease mechanisms and identify potential disease-causinggenes or associated biological pathways.Several studies have investigated comorbidity patterns of diseases byanalyzing the relevant biological factors[2-9]. Various mechanisms explain the etiology of comorbid diseases occurring in an individual. Shared disease genes are one of the biological factors responsible for disease comorbidities. Co-regulation of high-level biological mechanismssuch as the same cellular pathways can also contribute to the development of comorbid conditions. Interactions among proteins also led to disease comorbidities. The number of direct protein-protein interactions (PPIs) betweencausative proteins of two diseases has been considered to explainthe hidden comorbidity patterns. Recently, symptom similarity has also aidedthe understanding of unexpected association among diseases, diseaseetiology, and drug design.

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 research in disease comorbidity prediction [2] 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 cosine similarity, disease comorbidity patterns were generated.

Combined influence of various biological factors such as genes, pathways, protein-protein interactions, Gene ontology terms such as biological processes, cellular components and molecular functions had been considered in the proposed model. Consequently, these analyses helped usdiscover the associated disease mechanisms underlying comorbiddiseases and co-emerged clinical disease categories. Biological factors interleaved within direct and indirect relations were used to explore novel comorbidity patterns in asystematic manner.

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[10], HPRD[11] and IntAct[12]. CTD[13] and DisGeNet [14] 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 correspond 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 given in Fig.1. In the proposed method, dataset ranking was done using the dataset ranking algorithm published in our earlier work[15]. 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 [2] based on similarity measurement. Pathway dataset is ranked highest as per the decision analysis.

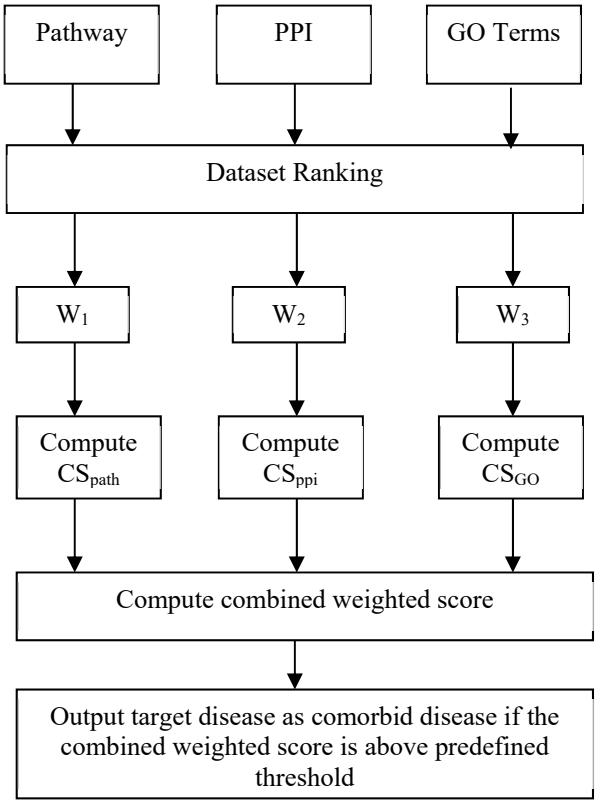

Fig.1. Overall System Architecture

The pseudo-code of dataset ranking is shown 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 $TP(D_i)$, $TN(D_i)$, $FN(D_i)$
For i in 1 to 3
For j in 1 to 3

$$C(D_i, D_j) = \frac{\sum_{\forall j: g(D_i) \geq g(D_j)} W_j}{\sum_{j=1}^{n} W_j}$$

If $\forall_x$, $g_x(D_i) \geq g_x(D_j)$ then
$D(D_i, D_j) = 0$
else

$$D(D_i, D_j) = \frac{1}{\delta} max\,[g_y(D_j) - g_y(D_i)]$$

If $C(D_i, D_j) \geq c_t$ and $D(D_i, D_j)) \leq d_t$ then D_i outperforms D_j;
Rank D_i accordingly with the highest weight w_x.

After finding the rank of each dataset, a weight was assigned to the dataset depending on the rank. Then the similarity between query disease and target disease was calculated using vector based approach. A total of 39239 PPI was considered. In the query disease vector, 1 was set corresponding to all PPIs which are associated with the given query disease and 0 for all other PPIs. Similarly, target disease vector was also created. Then the cosine similarity between two vectors was calculated using equation 1:

$$CS_{ppi} = \frac{\overrightarrow{QD_{ppi}} . \overrightarrow{TD_{ppi}}}{|\overrightarrow{QD_{ppi}}| . |\overrightarrow{TD_{ppi}}|} \qquad (1)$$

Cosine similarities between query disease and target disease were also calculated in terms of pathway (CS_{path}) and Gene Ontology terms (CS_{GO}) in the same way as done for PPI data. For finding pathway based similarity, disease-

pathway associations from CTD database was used. CTD contains phenotypic information as associations between terms in Gene Ontology and diseases. Three types of associations are available in CTD dataset: Biological Processes-Disease associations, Cellular Component-Disease associations, Molecular Functions-Disease associations. These associations were used for finding GO terms based similarity. Finally the comorbid score was generated using weighted sum (see equation 2).

$$WCS = \frac{W_1 CS_{ppi} + W_2 CS_{path} + W_3 CS_{GO}}{W_1 + W_2 + W_3} \qquad (2)$$

The proposed algorithm is given below:

Algorithm
Input: PPI Data, Pathway Data, GO Annotations, HPO Annotations
Steps:

1. Start
2. Input query disease, target disease, PPI data from PPI database, Disease-Pathway association and Disease - Gene Ontology (GO) term association from CTD database
3. Rank the dataset based on multi-criteria decision analysis using ELECTRE – I method.
4. Based on the rank, assign weights to PPI, Pathway and GO dataset as w_1, w_2 and w_3.
5. Initialize QueryDisease-ppi ($\overrightarrow{QD_{ppi}}$) and TargetDisease-ppi ($\overrightarrow{TD_{ppi}}$) vector with known disease-ppi interaction
6. Compute cosine similarity (CS_{ppi}) between $\overrightarrow{QD_{ppi}}$ and $\overrightarrow{TD_{ppi}}$
7. Initialize QueryDisease-pathway ($\overrightarrow{QD_{path}}$) and TargetDisease-pathway ($\overrightarrow{TD_{path}}$) vector with known disease-pathway interaction
8. Compute cosine similarity (CS_{path}) between $\overrightarrow{QD_{path}}$ and $\overrightarrow{TD_{path}}$
9. Initialize QueryDisease-GOTerm ($\overrightarrow{QD_{GO}}$) and TargetDisease-pathway ($\overrightarrow{TD_{GO}}$) vector with known disease-GOTerm interaction
10. Compute cosine similarity (CS_{GO}) between $\overrightarrow{QD_{GO}}$ and $\overrightarrow{TD_{GO}}$
11. Compute the weighted comorbidity score as $WCS = \frac{W_1 CS_{ppi} + W_2 CS_{path} + W_3 CS_{GO}}{W_1 + W_2 + W_3}$
12. If WCS > predefined threshold then target disease is comorbid with query disease else not comorbid
13. Stop

4. Results and Discussions

For evaluating the performance of the proposed method, a gold standard set of 15 diseases were chosen. The gold standard set of diseases is given in Table I. Table II shows the comorbidities related to diabetes mellitus generated from Gene Ontology terms. Table III shows the top comorbidities generated from PPI dataset and Table IV depicts the comorbidities generated from Pathway dataset.

LOOCV (Leave one out cross validation) was applied. The rules were validated in literature. As a case study, Alzheimer's disease was chosen specifically and detailed analysis of comorbidity of Alzheimer's disease with Schizophrenia is given in Fig.2. Performance comparison of vector based approach based on combined dataset was compared with vector based approach using individual dataset. Fig.3 shows the performance of the proposed approach.

SI No	Disease ID	Disease Name
1	MESH:D010051	Ovarian Cancer
2	MESH:D004827	Epilepsy
3	MESH:D006973	Hypertension
4	MESH:D008223	Lymphoma
5	MESH:D007037	Hypothyroidism
6	MESH:D003704	Dementia
7	MESH:D007938	Leukemia
8	MESH:D009203	Myocardial Infarction
9	MESH:D065646	Thyroid Carcinoma
10	MESH:D001249	Asthma
11	MESH:D009202	Cardiomyopathy
12	MESH:D009503	Neutropenia
13	MESH:D001172	Rheumatoid arthritis
14	MESH:D003324	Coronary artery disease
15	MESH:D003920	Diabetes Mellitus

Table I: Standard Disease Set

SI No	Comorbid Disease	PMID
1	Chronic kidney disease	25305751
2	Pancreatic Cyst	18470259
3	Hepatic Fibrosis	21350583
4	Myocardial Infarction	1512357
5	Macrocephaly	24138066
6	Alzheimers Disease	30542257
7	Hypertension	1568757
8	Polycystic Ovary Syndrome	22698921

Table II. Comorbidities associated with Diabetes Mellitus using GO terms

Disesae1	Disease2	PMID
MESH:D006973	MESH:D009202	8731103
MESH:D065646	MESH:D006980	26108596
MESH:D009202	MESH:D000417	4268057
MESH:D006528	MESH:C537136	2828214
MESH:D009203	MESH:D006949	23402469
MESH:D003920	MESH: D012559	29754122
MESH:D008303	MESH:D008881	18333513
MESH:D008831	MESH: D000544	21297427
MESH:D004827	MESH:D001764	26586031
MESH:D009103	MESH:D008180	29662683
MESH:D004535	MESH:D019283	2810267

Table III. Disease Comorbidities using PPI dataset

Disesae1_ID	Disease2_ID	PMID
MESH:D007037	MESH:D006111	30214742
MESH:D007938	MESH:D003920	22745776
MESH:C564421	MESH:D015228	23738826
MESH:D003704	MESH:D010300	18822028
MESH:D001172	MESH:D012507	6541252
MESH: D000544	MESH:D003704	22034442
MESH:D007037	MESH:D050171	29515631
MESH:D001249	MESH:D015658	4235227
MESH:D003324	MESH: D000544	3163092
MESH:D001714	MESH: D000544	26312426
MESH: D000544	MESH: D012559	26312426

Table IV. Disease Comorbidities using Pathway dataset

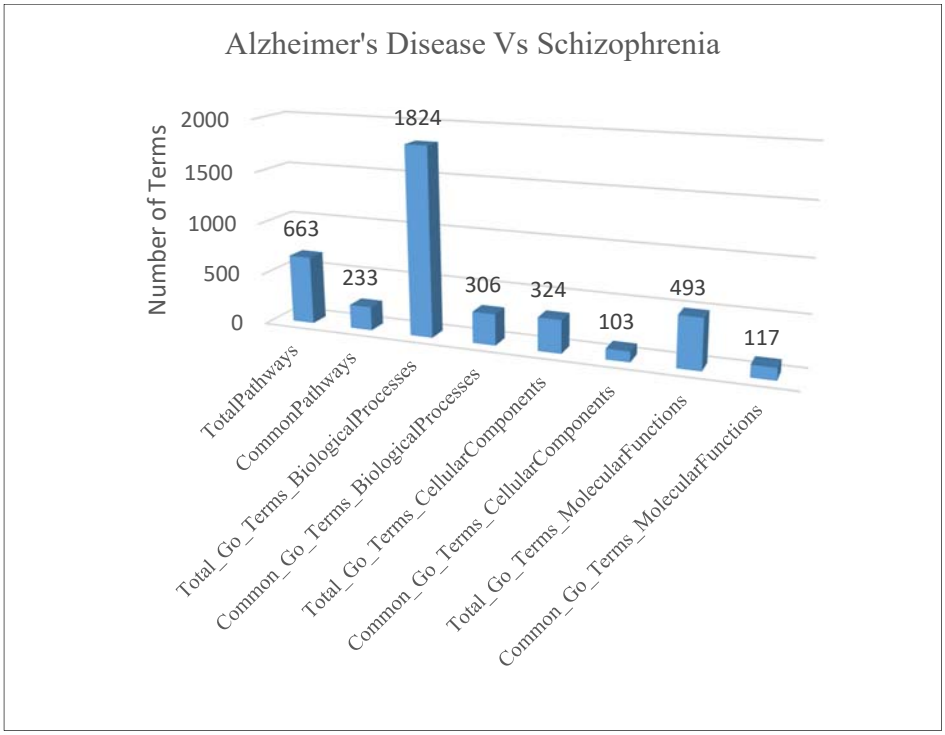

Fig.2.Comorbidity Statistics of Alzheimer's disease with Schizophrenia

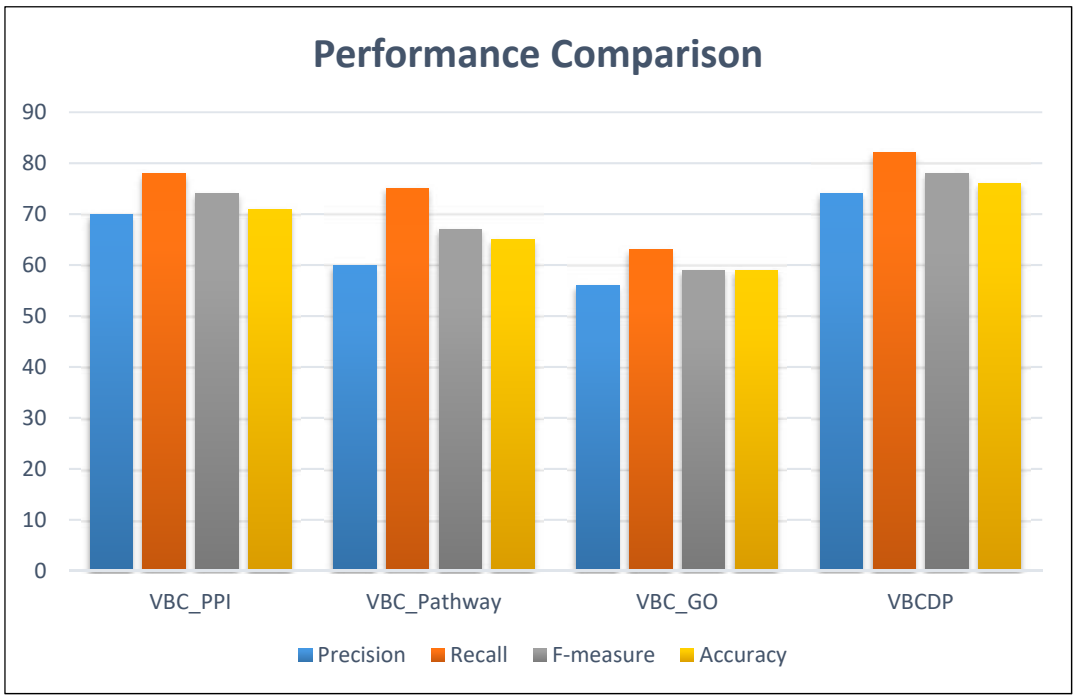

Fig.3. Performance Comparison

5. Conclusion

Discovering disease-disease associations and deep analysis of these findings help researchers to understand the basic underlying reasons behind development of such diseases and aids in the development of novel drugs. Malfunctioning of common cellular components, biological processes and metabolic functions in genes can lead to comorbid disease conditions. Doctors should be kept informed on novel information about the likelihood of different comorbid disease conditions.

The proposed approach 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 81.6%. The proposed approach 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.

References

[1] Roger Jones, "Chronic Disease and Comorbidity", British Journal of General Practice 2010; 60 (575): 394. doi:https://doi.org/10.3399/bjgp10X502056.
[2] Feng He, Guanghui Zhu, Yin-Ying Wang, Xing-Ming Zhao, De-Shuang Huang, "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, May/June 2017.
[3] Mohammad Ali Moni, PietroLiò, "comoR: a software for disease comorbidity risk assessment" Journal of Clinical Bioinformatics.2014.4:8. doi: 10.1186/2043-9113-4-8.
[4] Alba Gutierrez-Sacristan, Alex Bravo, Alexia Giannoula, Miguel A. Mayer, FerranSanz and Laura I. Furlong, "comoRbidity: an R package for the systematic analysis of disease comorbidities", Bioinformatics (2018) 1–3, doi: 10.1093/bioinformatics/bty315.
[5] YounheeKo, Minah Cho, Jin-Sung Lee, Jaebum Kim, "Identification of disease comorbidity through hidden molecular mechanisms" Scientific Reports 6:39433 (2016). https://doi.org/10.1038/srep3 9433.
[6] SachinMathur, DeendayalDinakarpandian, "Finding disease similarity based on implicit semantic similarity", Journal of Biomedical Informatics, Volume 45, Issue 2, April 2012, Pages 363-371. https://doi.org/10.1016/j.jbi.2011.11.017.
[7] Francesco Folino and Clara Pizzuti, "A Comorbidity-based Recommendation Engine for Disease Prediction" IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS). doi:10.1109/cbms.2010.6042664.
[8] Di Chen, Jin Tian, Yuepeng Yao, Songxing Du, JieyinGao, RongjuanGuo, Yun Wei, Peng Lu, "Recognition of Disease Comorbidity Medication Patterns Based on Network Motif Analysis" Research and Reviews: Journal of Pharmacy and Pharmaceutical Sciences (2016) Vol.5, Issue: 3.
[9] Khan A, Uddin S, Srinivasan U, "Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression", International Journal of Medical Informatics, Jul; 115:1-9. doi: 10.1016/j.ijmedinf.2018.04.001.
[10] Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012 Jan; 40(Database issue):D857-61. doi: 10.1093/nar/gkr930. Epub 2011 Nov 16.
[11] Prasad, T. S. K. et al. (2009). Human Protein Reference Database - 2009 Update. Nucleic Acids Research. 37, D767-72.
[12] Orchard S et.al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D358–D363, https://doi.org/10.1093/nar/gkt1115.
[13] Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, Wiegers TC, Mattingly CJ, "The Comparative Toxicogenomics Database: update 2019". Nucleic Acids Res. 2018 Sep 24.
[14] Janet Piñero, Juan Manuel Ramírez-Anguita, JosepSaüch-Pitarch, Francesco Ronzano, Emilio Centeno, FerranSanz, Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucl. Acids Res. (2019) doi:10.1093/nar/gkz1021.
[15] Lakshmi K.S, G.Vadivu, "A Novel Approach for Disease Comorbidity Prediction Using Weighted Association Rule Mining", Journal of Ambient Intelligence and Humanized Computing.

Ms 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 pursued her PhD from SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. Her research interest includes Data Mining and Bioinformatics. She has chosen Medical data mining as her research area. 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. She has more than 20 publications in reputed International Journals and has participated and presented papers in National and International conferences.

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 Transport Technology, 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, and now 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.