ISSN 2278-3091
Volume 8, No.4, July – August 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse12842019.pdf
https://doi.org/10.30534/ijatcse/2019/12842019

# Analysis of Human gene - Disease association as a Social network

**Ashwini R Doke[1], Nitin Garla[2], D. Radha[3],**
[1] Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India, ashwini.2688@yahoo.co.in
[2]Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India, nitin.1760@gmail.com
[3] Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India, d_radha@blr.amrita.edu

## ABSTRACT

Interrelationship of diseases can be analyzed using a biological network graph. In the Network graph of human diseases, each node represents a disease and interrelationship between the diseases is represented by as an edge. Interrelationship between the diseases represent the commonality in the genes associated with diseases. The size of a node represents the number of genes that are associated with the disease. Representation of such biological network can be used to analyze the way other social networks are analyzed. The proposed work introduces an approach to analyze biological network in terms of social network in which causes of different life-threatening diseases such as colon cancer are identified by using different centrality measures.

**Key words:** Biological network, Centrality measures. Colon cancer, Human diseases, Social network.

## 1. INTRODUCTION

A graph-based analysis of biological Network [1] plays an important role in the field of biomedical research. Although it is a complex task, it will achieve the goal of identifying strongly connected components in biological network which will be useful for treatment planning. Biological networks [2] consists of different components e.g. genes, protein, and metabolic etc. and to analyze these components by using graph representation makes biological network more feasible**.**

Biological network [3] is organized by different components of biology such as genetic network, protein network, Neuron network etc. that shows the interaction between genes, protein and neuron respectively. In all these networks certain nodes are abnormal in case of certain illness. These abnormal nodes may be cause for many other disorders. Identification of such nodes may help in the analysis of the cause of the disease.

Social network is study of individual, group, organization having similar interest like Twitter, Facebook etc. Such social network can be analyzed [4] by using different centrality measures as Closeness centrality, Eigen vector centrality, betweenness centrality and degree centrality to know the details about significance of the members in different aspects in that network. Degree centrality gives the density of the links a node has. Betweenness centrality is a measure which is based on shortest distance between two nodes. Closeness centrality gives the nodes which are closer to all other nodes. Eigen vector gives information about a node with significant neighbors.

The centrality measures used in social network [5] which identify the significance of nodes, can be used for the identification of critical nodes in biological network. In that sense the abnormal nodes that are significant for certain diseases can be identified.

A human disease [6] network which is a kind of biological network shows the relation between the diseases in terms of the genes associated with those disease. Centrality measures like degree centrality on HDN identifies a disease which may cause for many other diseases. betweenness refers to a disease which acts as an intermediator between other diseases. closeness centrality measure how a disease is directly or indirectly playing a role in another disease. Eigen-vector centrality shows the disease triggered by other critical diseases.

To know the significance of the nodes in the biological network, the measures that are used for a social network can also be applied in the biological network.

There are different tools has been used to analyze social network such as UCINET, GUESS, Pajek, Net Miner, Rob Cross, Gephi. Each tool has certain different features to analyse huge network. The same tools can be used to analyse biological networks.
There are different tools has been used to analyze social network such as UCINET, GUESS, Pajek, Net Miner, Rob

Cross, Gephi. Each tool has certain different features to analyse huge network. The same tools can be used to analyse biological networks.

In this paper an analysis of Human disease network is done by using Gephi tool. Gephi is graph visualization software for network analysis. It is an important and easy tool for analyzing and understanding large graphs. Manipulation of the colors, shapes, and structures of graph is done to reveal hidden patterns from the graph data. Gephi supports a wide range of graph formats. The relational databases and Comma Separated Values (CSV) can also be read using Gephi tool. Software updates for Gephi are automatically taken care by the built-in Plugins Center. There are dozens of community-built plugins that extends Gephi's functionalities. Diseasome is a Gephi dataset. Researchers used Gephi tool to create the diseasome graph which gives information about common genes associated for cause of different disorders.

In this paper biological network is analysed by using the Human disease network [7] (Diseasome Dataset) which is a network map of human diseases [8] and human genes and has been developed by the scientists to provide better visualization and analyses of relationship between diseases and genes associated with that disease.

## 2. LITERATURE REVIEW

Social network analysis [5] includes the study of relationship between social entities such as people, publications, web sites, nation states, organizations, or groups. Analysis and study of the social network is an important constituent of study of social science. Multiple statistical and mathematical tools which are used for analysis, are first developed in sociology [3]. There are multiple applications of social network analysis which includes analyzing the distribution of rumours, news and innovative ideas. Similarly, this approach can be used in medical [9] sector for analysis of health-related behaviours, issues and diseases.

Network analysis has also been applied in financial sector to study the markets. One of the advantages of using network analysis in financial sector is to identify easy mechanism for setting appropriate prices to achieve profit in business.

Social network analysis (SNA) can be used to study the recruitment into social organizations and political movements. Recently, SNA has been used to analyze the collective association among countries, institutions and authors in psychiatry research. Network analysis and traffic analysis has gained a lot attention in Military [33]. This approach is being used in military intelligence, to uncover rebellious networks of both categorized and leaderless nature e.g. terrorist organizations.

Researchers [10] have gained significant interest in analyzing molecular network due to easy availability of the biological data. This type of analysis is closely related to SNA, but local patterns in the network are mainly focused in such analysis. Advancement in the field of network medicine has been significantly increased by analyzing the biological networks with respect to diseases [11]. One of the applications of SNA in biology includes understanding the Cell Cycle. A physiological network can be seen as the communications between physiological systems like eyes, brain, heart etc.

Biological data include many biological components that makes a biggest challenge to analyse in biomedical field.

Research in graph-based approach describes how network representation can be useful for the researcher and clinician to better understand such complex data. Analysis can be performed with the help of various databases such as diseases database, protein interactions database [12], genetics database etc. using different computational tools. Diseasome, a network map of human diseases and human genes [13] has been developed by the scientists to provide better visualization and analyses the relationship between diseases and genes. Biological network [14] consist of data related to biology which includes genes, Protein, Metabolic related data [15]. To analyse such complex data by using graph representation make easy to understand the association between biological components.

Any biological systems can be used to form a biological network. Small units are linked together to form a network e.g. a food web network consists of different species and their linkage to each other. Relations found in physiological, evolutionary, and ecological studies can be mathematically represented with the help of biological networks. Medicine Network is the implementation of the concept of network science which include analysis of different human related disease in order to identify, prevent, and treat diseases like cancer [16].

Human genetics [17] includes study of genes associated with human body. That comprises of different fields such as biochemical genetics, genomics, population genetics, developmental genetics, clinical genetics, and genetic counselling. Human genetics study understands the human nature, as well as cause of disease. So, it will be helpful in treatment planning in medical domain.

In the era of big data, the importance of complex network [18] is an active area of research which includes study of social networks, technological networks. Study of Complex network in terms of a biological network [19] includes analysing biological data which consists of disorders and disease linked

together by common genes. This offers platform to identify common genetic origin of many diseases.

The review of literature describes the need of graph driven methods for analysis of biological data. The availability of biological data increases day by day which enables the need to systematically model the biological network in order to understand and analyze the behavior of biological components.

To identify the importance of biological systems, many scientists have used different approaches and methods that show common properties between different diseases. Different centrality measures are used for analysing such networks. Centrality [20] [21] measures are used to analyse the different communities in biological networks and to identify the relationship between genes and diseases. The analysis shows that when there is high betweenness centrality measure, then the interrelationship between them is high. Study [22] shows difference between social and biological network, carried out on protein interaction network. The most important node in protein network is identified by social network analysis measure. The same [23] can be used to determine protein function to identify proper drug for treatment. Weighted biological network [24] is also analyzed to study relationship between disease and genes.

The work [25] describes the network-based approach for human disease which identifies the gene associate with disease. The network is created with diseases as nodes and relationship between them are identified by considering different cellular components.

The analysis of network has been done by various methods as described in below section.There are many tools which have been used for graph analysis.

UCINET is a tool which is used for social network analysis. Net Draw is bundled with UCINET. It displays network visualizations and reading graphs is done with the help of NetDraw. UCINET is commonly used in academic world. Various visualization and other analysis platforms support the UCINET file format. There are some shortcomings with Ucinet e.g. filtering the data that is being viewed is difficult. Formatting of output is not allowed with NetDraw's visualization.

GUESS is a tool used to analyze and visualize graph and network data. This tool is widely used in academics. But installation of this tool is a little difficult as it has to be installed on a server. There is a lack of online documentation which helps in using the software. Advanced functionalities for formatting output is lacking in this tool. This tool requires additional time and more support for configuration.

Pajek tool is having a decent available documentation which helps to use the software. Pajek provides advanced functionality in the means of analyses of the data. Complex mappings of very enormous networks can be performed using Pajek.

Net Miner provides an intuitive and Convenient interface to user. This tool is helpful for the beginner and all other users who are less advanced. Its platform provides good support and documentation which makes it easy to use software online. Various visualizations and statistical analyses can be conducted using Net Miner. This tool is widely used by various users in different sectors such as general consumer use, corporate, and academic.

Rob Cross tool is advanced in terms of functionality but membership to Network Roundtable is required.
Gephi is open source software tool to analyze & visualize huge network. Motivation behind using this tool is, it provides support to data analytics to identify hidden patterns, making hypothesis, dividing structure uniqueness or faults during data sourcing. It is valuable for analyzing complex network. Gephi can read most graphs.

## 3. EXPERIMENTAL RESULTS

Analysis of biological network is prepared using human diseases network known as diseasome dataset. The tool used for the experiment is Gephi as it is an open source used for analyzing huge network in a visualized way.

Diseasome is a dataset of human disease network [26]. The network in Figure 1 is comprised of disease as nodes and edges represent the relationship between causes of two disease which has common genes. The size of a node indicates the genes associated with that disease. The disease nodes are connected to each other if the diseases are associated with common genes. The colored circular nodes in Figure 1 represent various disease classes. Shared genes among diseases are indicated using the lines connecting nodes. Thus, a colourful visual display using Gephi can be provided for diseasome network that helps the user to understand gene-disease [27] relationships in a better and simpler way.

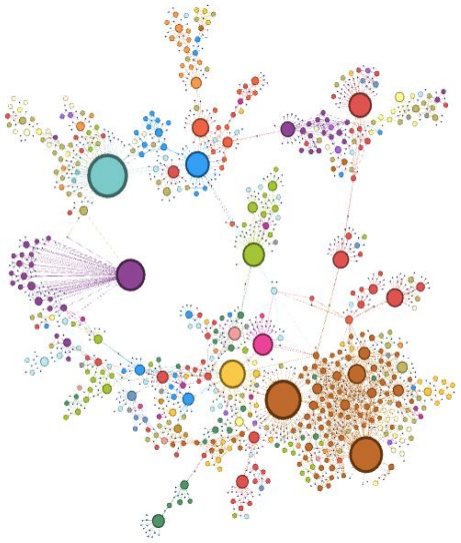

**Figure 1:** Diseasome Dataset

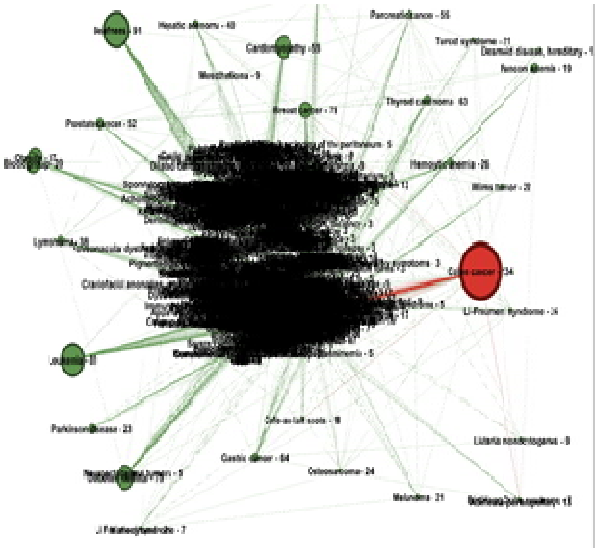

**Figure 2:** Highest degree centrality disease Colon cancer

The same can be used by biomedical researcher and health care solution providers for understanding genetic origin of many diseases. Analysis of the network [28], irrespective of its type, and size, can be performed using graphs. In this study, diseasome graph is analyzed and different centrality measures are calculated.

Centrality means a function in which each node in a network is assigned a numeric value C(v). The concept is used to rank the nodes and is helpful to identify important disease in the networks [29].

Degree centrality influences the importance of a particular node in that network. The importance of node will be higher if the number of neighbors of that node is more.

The Degree is calculated by the number of direct connections of each disease with all other diseases in the network. Degree centrality tells that how one disease is directly related with other diseases. Higher Degree centrality shows that number of diseases that directly caused by a disease or it can directly cause many other diseases. According to this in our measures of graph analysis shows, Colon cancer is having highest degree centrality which means that many other diseases like breast cancer, gastric cancer, leukemia may directly lead to colon cancer [30]. The same is verified from the biological disease analysis [26]. Figure 2 shows highest degree centrality disease i.e. Colon cancer.

The indegree centrality shows the number of ties that directs to a node. Indegree gives the count of total number of edges coming to a node.

According to human disease network, indegree centrality of a disease tells that the number of diseases that are triggered directly by a particular disease.

Figure 3 shows the highest In-degree centrality disease. Colon cancer is having highest indegree i.e. 84. Colon cancer may directly cause other diseases.

The outdegree centrality shows the number of ties that the node directs to others. Outdegree gives the count of total number of edges going out of a node.
According to human disease network, outdegree centrality provides the number of diseases that are directly caused by a disease.

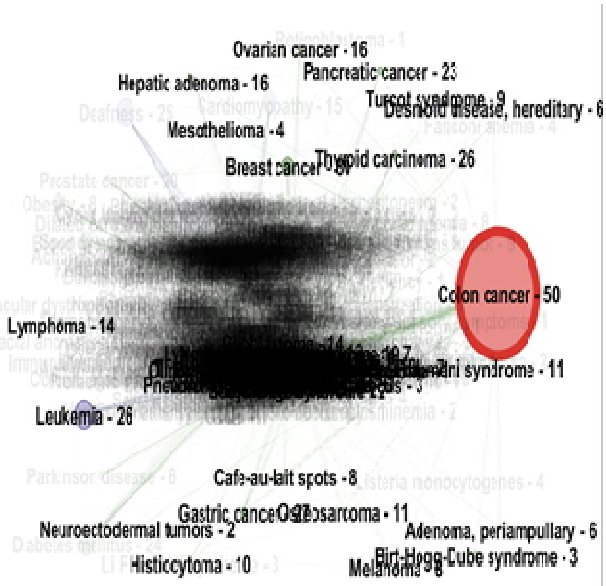


**Figure 3:** Highest In degree centrality disease Colon cancer

Figure 4 shows the highest Out-degree disease. Colon cancer is having highest out degree i.e. 50. Many other diseases are caused directly by Colon cancer.

Betweenness centrality measures the control of a node over the information flow in the network. Between ness centrality of any node A can be calculated by sum of total number of shortest paths between any two nodes which passes through node A divided by the total number of shortest paths between two nodes.

According to human disease network, betweenness refers to a disease which acts as an intermediator between any two other diseases. According to human disease network graph, if a person suffering from deafness and lipodystrophy syndrome then he is suffering from Cardiomyopathy because Cardiomyopathy acts as a bridge between deafness and Lipodystrophy. Cardiomyopathy is a cardio muscular disease and lipodystrophy is a metabolic [15] disease. The betweenness centrality proves that if the person having lipodystrophy and diabetes, must have cardiomyopathy which is discussed in the paper [15].

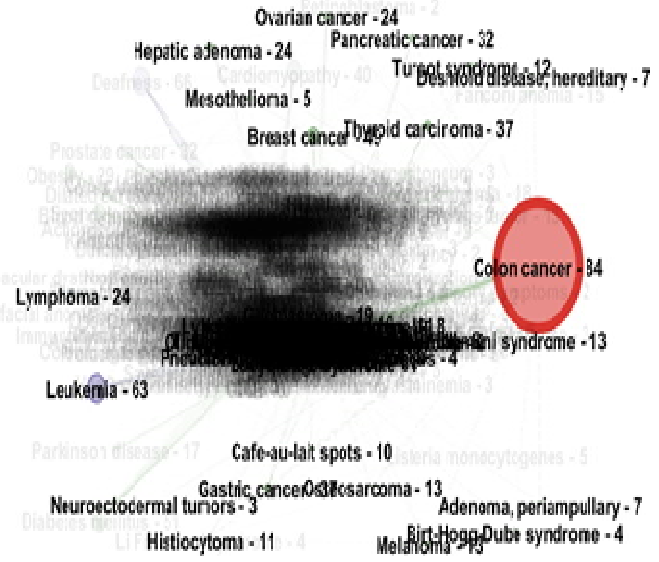


**Figure 4:** Highest Out-degree centrality disease Colon Cancer

Fig. 5 shows the highest betweenness centrality disease. The cardiomyopathy is having highest betweenness centrality which says that this disease related with many diseases.

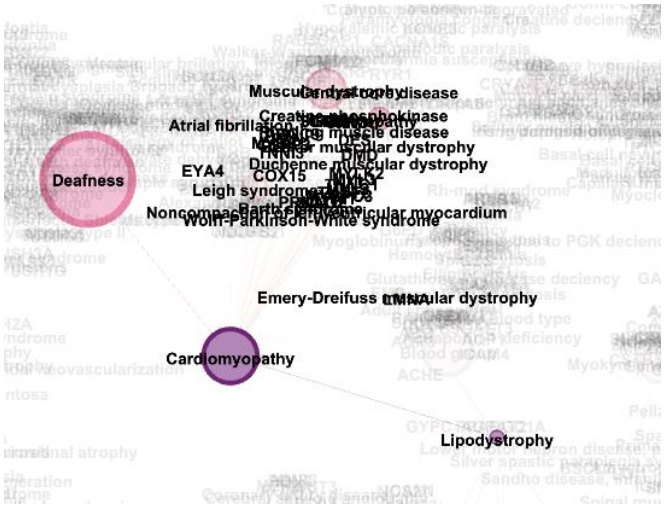


**Figure 5:** Highest Betweenness disease Cardiomyopathy

The other diseases having highest Betweenness centrality are lipodystrophy, diabetes mellitus, glioblastoma, deafness, myopathy, cataract, leukemia, colon cancer and #Alzheimer disease.

The closeness centrality of a node indicates how close the node from another node in a network. If the distance is lesser to other nodes, then it can be said that the node is closer to those nodes. Closeness is a measure of the degree to which an individual is near to all other individuals in a network. According to human disease network, closeness centrality measure notifies how a disease is directly or indirectly playing a role in another disease. Higher closeness centrality means the disease is directly or indirectly either caused by another disease or may lead to another disease.

Figure 6 shows the highest closeness centrality disease. In our result it has been shown that the Lipodystrophy is having highest closeness centrality.

**Figure 6:** Highest closeness centrality disease Lipodystrophy

As Lipodystrophy is co related with loss of body fat which may leads to diabetic nephropathy and retinopathy. This highest closeness centrality of Lipodystrophy predicts that it

may cause several other disorders.

Eigen vector Centrality measure is used to show the influence of a node in the any network based on its connections to other important nodes.

Colon cancer is having highest eigen vector centrality in Diseasome network. Colon cancer is triggered by other related diseases like breast cancer, blood cancer, gastric cancer, prostate cancer which in turn are triggered by many diseases. It is associated with more genes i.e. these diseases are playing important disease in human disease network and are to be well concentrated.

The Diseasome network is analysed with social network measures using the tool Gephi for the relationship between the diseases and genes associated with them.

**Table 1:** Centrality Measures of Diseasome network

| Measures used for analysis | Node with highest measure | Values |
|---|---|---|
| In-degree | Colon cancer—disease | 50 |
| out-degree | Colon cancer—disease | 84 |
| Degree | Colon cancer—disease | 134 |
| Closeness Centrality | Lipodystrophy—disease | 0.245414 |
| Betweenness Centrality | Cardiomyopathy —disease | 0.166189 |
| Eigen Vector Centrality | Colon cancer— disease | 1.0 |

As shown in table one the disease colon cancer has highest measure of indegree, outdegree and degree centrality. That is colon cancer may lead to many diseases.as it contains the genes which are common to many diseases. Figure 7 shows the graphical representation of diseases(node) having highest centrality measure. From graph it can be analysed that colon cancer is having highest centrality measure, so it is an important disease which will cause many diseases.

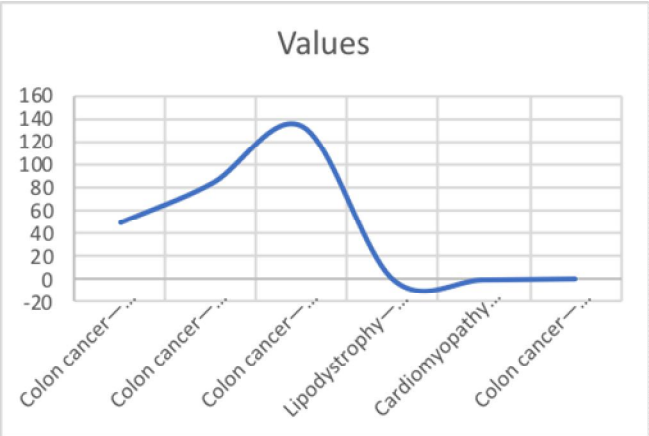


**Figure 7:** Analysis of Human disease network

## 5. CONCLUSION

These analysis helps in finding the common genetic origin of many diseases which will be helpful to identify the causes of different disease and to make proper drug design and treatment plan. Disease and all the genes associated with that disease can be analyzed with the help of diseasome graph. This provides a better visual genetic links between disease gene and disorder which will be useful for physician, genetic counselors and biomedical researchers. Minor diseases can be analysed and find the causes for Dangerous disease which helps in solving the problem in the earlier stage.

## REFERENCES

1. Hakes, L., Pinney, J.W., Robertson, D.L. and Lovell, S.C., 2008. **Protein-protein interaction networks and biology— what's the connection?** In Nature biotechnology, 26(1), p.69. https://doi.org/10.1038/nbt0108-69
2. Euiseong Ko, Mingon Kang, Hyung Jae Chang, Donghyun Kim. **Graph-Theory Based Simplification Techniques for Efficient Biological Network Analysis,** IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), pp. 277 – 280.
3. Newman, M.E.J. **Networks: An Introduction**. Oxford University Press. 20
4. Grando, F. **On the analysis of centrality measures for complex and social networks**, 2015. https://doi.org/10.1109/GLOCOM.2016.7841580
5. Wang, P., Xu, B., Wu, Y. and Zhou, X. **Link prediction in social networks: the state-of-the-art.** In: Science China Information Sciences, 58(1), pp.1-38 (2017). https://doi.org/10.1007/s11432-014-5237-y
6. Goh, K.I. and Choi, I.G., 2012. **Exploring the human diseasome: the human disease network**. In: Briefings in functional genomics, 11(6), pp.533-542 (2012). https://doi.org/10.1093/bfgp/els032
7. Jimenez-Sanchez, G., Childs, B. and Valle, D. **Human disease genes.** In Nature, 409(6822), p.853 (2001). https://doi.org/10.1038/35057050
8. Zhu, X., Gerstein, M. and Snyder, M. **Getting connected: analysis and principles of biological networks**. In Genes & development, 21(9), pp.1010-1024 (2007). https://doi.org/10.1101/gad.1528707
9. Barabási, A.L. **Network medicine—from obesity to the "diseasome"**. In: N. Engl. J. Med. 357, 404–407 (2007). https://doi.org/10.1056/NEJMe078114
10. Helikar, T., Kochi, N., Konvalina, J. and Rogers, J.A. **Boolean modeling of biochemical networks**. In The Open Bioinformatics Journal, 5, pp.16-25 (2011). https://doi.org/10.2174/1875036201105010016
11. Barabási, A.L., Gulbahce, N. and Loscalzo, J. **Network medicine: a network-based approach to human disease**. In Nature reviews genetics, 12(1), p.56 (2011).

https://doi.org/10.1038/nrg2918

12. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S. and Timm, J. **A human protein-protein interaction network: a resource for annotating the proteome**. In Cell, 122(6), pp.957-968 (2005). https://doi.org/10.1016/j.cell.2005.08.029
13. Erkan, G. and Radev, D.R.: Lexrank: **Graph-based lexical centrality as salience in text summarization**. In Journal of artificial intelligence research, 22, pp.457-479 (2004). https://doi.org/10.1613/jair.1523
14. Özgür, A., Vu, T., Erkan, G. and Radev, D.R.: **Identifying gene-disease associations using centrality on a literature mined gene-interaction network**. In Bioinformatics, 24(13), pp. i277-i285 (2008). https://doi.org/10.1093/bioinformatics/btn182
15. Hadigan, C., Meigs, J.B., Corcoran, C., Rietschel, P., Piecuch, S., Basgoz, N., Davis, B., Sax, P., Stanley, T., Wilson, P.W. and D'agostino, R.B.: **Metabolic abnormalities and cardiovascular disease risk factors in adults with human immunodeficiency virus infection and lipodystrophy**. In Clinical Infectious Diseases, 32(1), pp.130-139 (2001). https://doi.org/10.1086/317541
16. Ergün, A., Lawrence, C.A., Kohanski, M.A., Brennan, T.A. and Collins, J.J. **A network biology approach to prostate cancer.** In Molecular systems biology, 3(1), p.82 (2007). https://doi.org/10.1038/msb4100125
17. Midic, U., Oldfield, C.J., Dunker, A.K., Obradovic, Z. and Uversky, V.N. **Unfoldomics of human genetic diseases: illustrative examples of ordered and intrinsically disordered members of the human diseasome**. In: Protein and peptide letters, 16(12), pp.1533-1547 (2009). https://doi.org/10.2174/092986609789839377
18. Metlicka, M. and Davendra, D.: **Complex network analysis based adaptive differential evolution**, IEEE Congress on Evolutionary Computation (CEC) pp. 3338-3345 (2016). https://doi.org/10.1109/CEC.2016.7744212
19. Zhu, X., Gerstein, M. and Snyder, M.: **Getting connected: analysis and principles of biological networks**. In Genes & development, 21(9), pp.1010 -1024 (2007). https://doi.org/10.1101/gad.1528707
20. Radha, D., Kavikuil, K. and Keerthi, R.: **Centrality Measures to Analyze Transport Network for Congestion Free Shipment**, 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), pp. 1-5 (2017). https://doi.org/10.1101/gad.1528707
21. A. Adamic, Lada & Wilkinson, Dennis & Huberman, Bernardo & Adar, Eytan. **A literature based method for identifying gene-disease connections**, IEEE Computer Society Bioinformatics Conference, pp. 109-17.
22. Zhang, X.Y., Wang, X.H. and Huang, C. **Research on the knowledge sharing of the university interdisciplinary team based on social network analysis**, International Conference on Management Science and Engineering (ICMSE), pp. 478-485 (2016). https://doi.org/10.1109/ICMSE.2016.8365476
23. Radha, D. and Kulkarni, S. **A Social Network Analysis of World Cities Network**, 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), pp. 1-6 (2016). https://doi.org/10.1109/CSITSS.2017.8447571
24. Sibbald, S.L., Wathen, C.N., Kothari, A. and Day, A.M. **Knowledge flow and exchange in interdisciplinary primary health care teams (PHCTs): an exploratory study**. In Journal of the Medical Library Association: JMLA, 101(2), p.128 (2013). https://doi.org/10.3163/1536-5050.101.2.008
25. Ding, D. **On the centrality of complex human disease network**. In 30th Chinese Control Conference (CCC), pp. 5627-5629 (2011).
26. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. **The human disease network**. In Proceedings of the National Academy of Sciences, 104(21), pp. 8685-90 (2007). https://doi.org/10.1073/pnas.0701361104
27. Hu, G. and Agarwal, P. **Human disease-drug network based on genomic expression profiles**. In PloS one, 4(8), pp. 6536 (2009). https://doi.org/10.1371/journal.pone.0006536
28. Yıldırım, Muhammed A., Kwang-Il Goh, Michael E. Cusick, Albert-László Barabási, and Marc Vidal. **Drug—target network**. In Nature biotechnology 25, no. 10, pp. 1119 (2007). https://doi.org/10.1038/nbt1338
29. Malathi, A. and Radha, D. **Analysis and visualization of social media networks**, International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), pp. 58-63 (2016). https://doi.org/10.1109/CSITSS.2016.7779440
30. Cai, Q.G.Y.T., Chow, W.H.S.X.O., Yang, G.J.B.T., Wen, W.R.N., Li, H.L.M.J.D. and Zheng, W. **Prospective study of urinary prostaglandin E2 metabolite and colorectal cancer risk,** In Journal of Clinical Oncology, 24(31), pp.5010-5016 (2006). https://doi.org/10.1200/JCO.2006.06.4931