www.revistageintec.net
ISSN: 2237-0722

revista GEINTEC
Gestão, Inovação e Tecnologias

Identifying Disease Comorbidity Patterns Using Centrality Measures in Computing

Zahra Batool[1]; Muhammad Junaid[2]; Muhammad Naeem[3]; Mehmood Ahmed[4]; Luqman Shah[5]; Yousaf Saeed[6]; Ali Imran Jehangiri[7]; Fahad Ali Khan[8]

[1]Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Pakistan.
[2]Department of IT, Abbottabad University of Sciences and Technology, Abbottabad, Pakistan.
[2]mjunaid@uoh.edu.pk
[3]Department of IT, Abbottabad University of Sciences and Technology, Abbottabad, Pakistan.
[4]Department of IT, The University of Haripur, Pakistan.
[5]Department of IT, The University of Haripur, Pakistan.
[6]Department of IT, The University of Haripur, Pakistan.
[7]Department of IT, Hazara University, Mansehra, Pakistan.
[8]Department of IT, Abbottabad University of Sciences and Technology, Abbottabad, Pakistan.

Abstract

Social network analysis has been increasingly employed to study patterns in diverse areas of disciplines such as crowd management, air passenger and freight transportation, business modelling and analysis, online social movements and bioinformatics. Over the years, human disease networks have been studied to analyze Human Disease, Genotype, and Phenotype networks. This study explores human Disease Network based on their symptoms by employing different social network analysis such as centrality measures of network, community detection, overlapping communities. We studied relationships of symptoms with diseases on meso-level in order to detect comorbidity pattern of communities in disease network. This help us to understand the underlying patterns of diseases based on symptoms and find out that how different disease communities are correlated by detecting overlapping communities.

Key-words: Network Analysis, Disease Network, Disease Association, Centrality Measures, Community Detection, Meso Level.

1. Introduction

Over few decades there has been a lot of study on finding diseases on the basis of their related factors including social biological factors, and their underlying genetic components [1,10]. Different diseases are correlated, if someone get suffered from one disease then it is likely to suffer from others as well. Many human diseases are related to each other through shared causes, shared genes, shared PPIs or even shared pathology [13]. Over the course of recent years, a significant portion of bioinformatics research has been focused on this aspect [1,14]. Without knowing how different diseases are connected to one another, our understanding of human diseases is inadequate. Disease relationships have been utilised to learn more about the origin and pathophysiology of related diseases [5,10].

A number of resources have been constructed aims to understand the entangled relationship between diseases, hidden in complex disease networks. But most of the studies focused on micro and macro aspects of human disease network, meso-level does not studied in depth yet. In context of social networks, micro level begins with an individual or may begin with a small group of individuals. Macro-level analyses generally trace the outcomes of interactions such as economic and meso-level falls between the micro and macro-levels and focus on the community structure of a network to study how biological networks can evolve over time. Diseases are interconnected based on their related symptoms. There may be a reason that disease is the result of previous exposure to other diseases. According to research, 99% of people who have diabetes are likely to discover cardiovascular disease [15]. This disease etiology help medical researchers to find root cause of disease.

In previous studies [1,10] micro and macro aspects are discussed, while we investigated the human symptoms-disease network at meso-level. By exploiting the dynamics of network analysis, we may attempted to predict the future of disease clusters. In this paper, we focus on detecting diseases propagation and comorbidity patterns and try to figure out that how different disease communities are inter-related. We also intend to study that how these diseases are spread and propagates in a network. This study has, therefore, employed to study disease comorbidity pattern based on their underlying symptoms. We have studied disease association patterns using mayo clinic data consisting of 644 diseases and 68 symptoms. In our disease symptoms network, firstly we have generated a bipartite graph from edge list, where the link weight between two diseases quantifies the similarity of their respective symptoms, then we constructed our network using edge list, which quantifies the relation between symptoms and diseases. Degree, betweenness and closeness centrality measures are used to study most dominating diseases in a network. Finally, the Community detection methods are used to group diseases based on their underlying symptoms.

Received: 08.06.2021 – Accepted: 09.07.2021

The rest of the paper is organized as follows Section 2 discusses the related work with their limitations. Section 3 includes proposed methodology, Section 4 discusses results and analysis of the study. Finally in section 5, we present the conclusions and future work.

2. Related Work

The analysis of disease network is performed by using symptoms data, disease-genes PPIs data, and by using patient data. Kim et al. constructed a human disease network by using claims data and investigated disease–disease associations from different perspectives by conducting network analysis [3]. They also explored diseases which are associated with the major causes of mortality and morbidity. Zhou, et al. built a symptom-based human disease network to find groups associations between symptom similarity of diseases, shared genes and PPI's [1]. Menche et al. proposed a network-based approach for locating illness modules within the interaction and predicting disease-disease correlations based on overlap between modules [14]. Disease pairings that were projected to have overlapping modules showed statistically substantial molecular similarity, according to the network-based distance between two disease modules. Sun et al. studied network by examining computationally predicted disease associations [2]. In 2007, Goh et al. built a bipartite graph of diseases connected with their associated genes [4]. The researchers investigated illness connections by building a human disease network, in which two diseases are connected if they share at least one gene. Moreover, in 2012, researchers studied how the connectivity between molecular parts translates into the relationships between the related diseases [5]. In 2009, Li et al. investigated disease relationships based on their shared pathways [10]. The researchers created a disease network by tying illnesses together when they have similar pathways. Vanunu et al. employed a protein network to extract a neighbourhood of candidate proteins for the design of a prioritisation function to predict gene-disease associations [6]. Lee et al. presented a metabolic disease network, by constructing a bipartite human disease association network in which nodes are diseases and two diseases are linked if mutated enzymes associated with them catalyze adjacent metabolic reactions [7]. Liu et al. combined the environmental etiological factors and genetic factors to expose similarities between diseases [8]. Roque et al. used electronic patient record to reveal comorbidity patterns among diseases [11]. Lee et al, constructed a disease network to study association among diseases using PPIs, EHR, biological pathways and biomedical literature data [12].

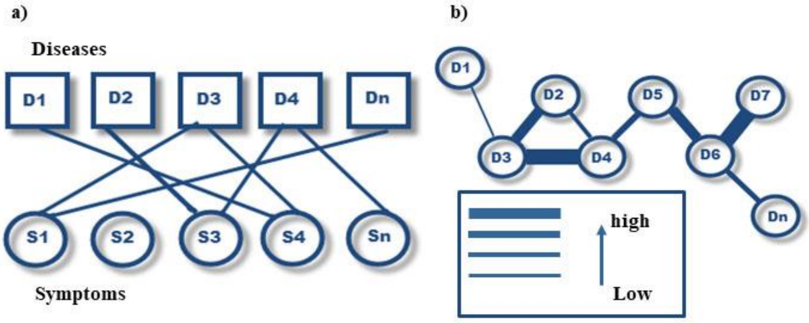
Many human diseases are linked by common causes or even pathophysiology. Similar illnesses have traditionally been treated using knowledge of these connections [9]. By reviewing previous studies related to finding disease comorbidity patterns, we observed that there are many studies focus on detecting disease-disease relationships on basis of their symptoms similarity, gene similarity, protein to protein interaction, and biological pathways, which covered micro, macro level analysis. The meso level network study can be a useful in finding disease association by finding overlapping communities in a network.

3. Proposed Methadology

In this analysis, we collected data from http://www.mayoclinic.org/ for generating a disease symptom network. Data consists of two columns, disease and symptoms. We have extracted total of **5977** records after cleaning data, we delete duplicated rows, and only **1857** records were left. After filtering for the co-occurrence of at least one disease and one symptom total **644** diseases and **68** symptoms left. Starting from the disease-symptom bipartite graph, we have obtained disease-to-disease network projection. The Human Disease Network (HDN), nodes represent diseases, and two diseases are connected with each other if they share at least one symptom. The bipartite graph representation and their resulted disease to disease network projection is shown in figure 1.

We have extracted weighted edge list of diseases. This edge list have total number of **62188** edges and **644** nodes. Nodes shows the diseases and edges shows the symptoms in our disease network. Then, we analyzed our data by using cytoscape, GGally, igraph and Linkcomm library in R which used clustering of communities structure of the network (Ahn et al., 2010; Evans and Lambiotte, 2009).

Fig. 1 - a) The Association between Disease and Relationship Extracted from Mayoclinic Data. b) Disease Network Constructed based on their Underlying Symptoms, Weighted Edges shows Intensity of Disease Association based on their Symptoms



We implemented the algorithm outlined by Ahn et al. (2010), which employs the Jaccard coefficient for assigning similarity between links,eik and ejk , that share a node k.

$$S(eik,ejk) = |ni \cap nj| / |ni \cup nj|$$

where n+(i) refers to the first-order node neighborhood of node i. The linkages are hierarchically grouped after pairwise similarities are assigned to all of the links in the network, and the resultant dendogram is sliced at a point that optimises the density of links inside the clusters. Normalizing against the partition density, which is the maximum and lowest number of connections in each cluster. Using single linkage, we conducted hierarchical clustering. A pictorial view of the methodology, designed for this study, is depicted in figure 2.

Fig. 2 - Proposed Methodology

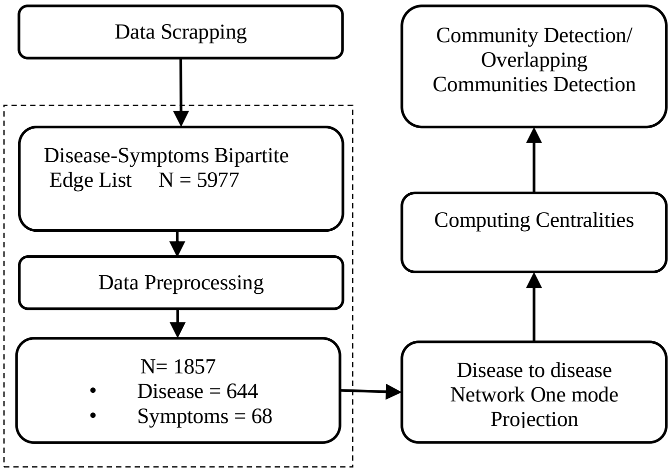

Data Scrapping
Disease-Symptoms Bipartite Edge List N = 5977
Data Preprocessing
N= 1857
• Disease = 644
• Symptoms = 68
Disease to disease Network One mode Projection
Computing Centralities
Community Detection/ Overlapping Communities Detection

Table I - Micro Level Statistics

Metrics	**Value**
Radius of the network	2
Average Path length	1.76
Number of shortest paths	410240
Average Clustering Coefficient	0.804
Total triangles in the networks	3978505
Average Degree	193.130
Largest clique	196
Average number of neighbors	192.894.

4. Results and Analysis

The Disease Network is very dense having density 0.300. The shortest distance between two disease nodes i.e. diameter of a network is 4, shows connectivity of a network. It means we need 4 steps to reach other disease node in a network. Table 1 gives the micro level statistics related to the disease to disease network.

Centrality Measures

Centrality measures are common method that is used for identifying mostly connected diseases in disease network. Diseases which are common and mostly linked with each other, can be identified using centrality measures. We used four common centrality measures in our analysis i.e. degree, closeness, betweenness, and eigenvector centrality.

Degree Centrality: Nodes having high degree centrality are the most connected nodes in the network. The degree centrality is represented as

$$C_D\,(n_i) = d(n_i)$$

where *d(n_i)* is the degree of node n_i. The degree based centrality finds the node which has the most connection with other nodes. This aspect highlights the diseases that linked with other diseases, which are helpful in finding disease correlation. Table 2 shows the list of top 20 diseases that we ranked according to degree centrality measures in disease network. These diseases are the most interlinked with other in the network, based on their underlying symptoms, like fever, flue, and cough. Hyperglycemia in diabetes and Churg-Strauss syndrome have common symptoms i.e. Shortness of breath, Fatigue, Nausea and vomiting, and Abdominal pain. Fatigue, Nausea and vomiting, Headache, Diarrhea, unexplained weight loss and cough are the most co-exist symptoms in the diseases listed in Table 2.

Closenes Centrality: Closeness is a measure of how long it takes the information to spread from a given vertex to others in a network [16]. Closeness can be measured by

$$C_c\,(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$$

where *d (i, j)* is the distance between two vertices in a network. In closeness centrality, we calculated the node centrality w.r.t. the average distance of a node to all other nodes.

Table II - Degree and Closeness Centrality

Disease Name	**Degree**	**Closeness**
Hyperglycemia in diabetes	450	0.771084
Churg-Strauss syndrome	446	0.766467
Legionnaires' disease	443	0.758294
Lead poisoning	439,	0.759193
Polio	435	0.756501
Chronic sinusitis	432	0.748538
Hemophilia	429	0.750283
Angina	427	0.746791
Ascariasis	425	0.746791
Plague	425	0.745921

Received: 08.06.2021 – Accepted: 09.07.2021

The disease Hyperglycemia in diabetes, Churg-Strauss syndrome, Lead poisoning, Legionnaires disease, Polio, and Hemophilia have highest closeness, as shown in Table 2. It depicts that given nodes are the most connected nearest nodes to others in a network. In contrast, Genital warts, Sprained ankle, Low sperm count, Male infertility, Spermatocele, esticular cancer, and Bags under eyes have the lowest closeness in the network. It is shown that these diseases are not common and not linked to other diseases based on their underlying symptoms.

Betweenness Centrality: Betweenness centrality shows the number of shortest paths passing through a vertex. High betweenness shows the number of largest connecting groups to a node, meaning that a node act as bridge between nodes and monitors the flow of knowledge between the other nodes. The betweenness centrality, C_B***, determines the shortest path among two nodes in a network [1] and is defined as***

$$C_B(m_i) = \sum_{j<k} h_{jk}(m_i) / h_{jk}$$

where $h_{jk}(m_i)$ is the number of shortest paths from nodes j to k through node i, and h_{jk} is the number of shortest paths connecting two nodes *j* and *k*. Betweenness centrality ranks the disease nodes with highest value that have most shortest path to other nodes. The disease named ‘Hemophilia’ has the highest betweenness as Hemophilia has Elbow pain, Fatigue, Nausea and vomiting, Headache, Hip pain, Joint pain, Knee pain, Muscle pain, Neck pain, and Nosebleeds like common symptoms which also exist in most of other disease. Diseases with highest betweenness centrality are shown in Table 2.

Eigenvector centrality: Based on the idea that a node is more central if it is in relation with other nodes. It is computed as

$$x_i = \frac{1}{\gamma} \sum_{j \epsilon M(i)} x_j$$

where $j \epsilon M(i)$ means that the sum is over all *j* such that the nodes *i,j* are connected. With increasing degree, eigenvector value is also increasing. The diseases with highest eigenvector centrality are given in Table 4, which shows that the Hyperglycemia in diabetes and Legionnaires diseases are the most influential nodes in the disease network w.r.t eigenvector centrality. The most dominant diseases in the network are listed in Table 4.

Community Detection and Overlapping Community Detection

The modularity of network is 0.201 which shows that this network has good community structure and total 5 communities found in disease network.

Table III - Betweenness Centrality

Disease Name	Betweenness
Hemophilia	5072.300846
Recurrent breast cancer	4452.720424
Fibromuscular dysplasia	2807.034604
Pelvic inflammatory disease (PID)	2525.287496
Lyme disease	2416.839923
Gonorrhea	2409.631343
Diabetic neuropathy	2209.711618
Porphyria	2115.428126
Behcet's disease	2071.175316
Polymyalgia rheumatica	1932.814829

Table IV - Eigenvector Centrality

Disease Name	Eigenvector
Hyperglycemia in diabetes	1
Legionnaires' disease	0.992234
Churg-Strauss syndrome	0.982835
Chronic sinusitis	0.978347
Lead poisoning	0.974585
Plague	0.971039
Ascariasis	0.968868
Chagas disease	0.963682
Q fever	0.963561
Swine flu (H1N1 flu)	0.963561

Spencer circle layout of disease network shown in figure 3. This layout dispersed communities evenly around the circumference of a circle in dendrogram order (to reduce link crossing over) and placed nodes within the circle according to the number of links they have in each of the communities. As a result, nodes with many connections are pushed to the centre of the circle. The figure 3 shows that each node is depicted as pie chart, illustrating its membership in multiple communities. The Largest cluster consist of 31.36% nodes of the network includes diseases like angina, Aortic aneurysm, Aortic dissection, Chronic lymphocytic leukemia, Churg-Strauss syndrome, Enlarged spleen, Kidney cancer, Mesothelioma, Non-Hodgkin's lymphoma, Nonalcoholic fatty liver disease, Polymyalgia rheumatica, Small vessel disease, Tuberculosis, Vasculitis and many more.

Abdominal pain, back pain, cough, shortness of breath, fatigue, and unexplained weight loss are the most common symptoms that connect these diseases to form a cluster. Acromegaly, Behcet's disease, Bone spurs, Brucellosis, Dengue fever, Depression, Eyestrain, Fibromuscular dysplasia, Influenza (flu), Lead poisoning, Mental illness, Pheochromocytoma, Polio and others diseases form another cluster consists of 29.02% network, these disease are mostly related based on symptoms like headache, fatigue, back pain. Third cluster representing 27.15% of network, consist of Abdominal aortic aneurysm, Acute liver failure, Addison's disease, Alcoholic hepatitis, Anaphylaxis, Antibiotic-associated diarrhea, Appendicitis, Ascariasis, Barrett's esophagus, Bile reflux, Bladder stones, Blastocystis hominis infection, Blind loop syndrome, Botulism, C. difficile, Carcinoid tumors, Cardiogenic shock, Celiac disease, Chagas disease, and others. These diseases are connected based on symptoms i.e. abdominal pain, Diarrhea and nausea, and unexpected weight loss. Arthroscopy, Avascular necrosis, Bacterial vaginosis, Baker's cyst, Benign prostatic hyperplasia (BPH), Bladder cancer, Buerger's disease, Cervical cancer, Cervicitis, Cystitis, Diabetic hyperosmolar syndrome, Endometrial cancer, Flatfeet, Genital warts, Gestational diabetes, Growing pains, and Guillain-Barre syndrome form another cluster which illustrated only 11.23% of network. Leg pain, pelvic pain, frequent urination, vaginal bleeding are the most common symptoms in this network. The smaller cluster covers diseases i.e. Breast cancer, Ductal carcinoma in situ (DCIS), Fibrocystic breasts, Male breast cancer, Mammary duct ectasia, Paget's disease of the breast, Breast cysts. These diseases are connected to form smallest cluster in the network with only 1.25% of nodes. Moreover, symptoms i.e. Breast lumps, Nipple discharge are the only two underlying symptoms in this cluster, which depicts that this diseases targeted only specific diseases related to anatomical localization of human body.

Figure 4 shows a community membership matrix with color-coded community membership for nodes that belong to the most communities. We were able to extract meta communities by grouping these communities further. Nodes may appear numerous times across various meta-communities. Churg-Strauss syndrome, HIV/AIDS, Legionnaires' disease, Ascariasis, Wegener's granulomatosis, Aspergillosis, Celiac disease, Hantavirus pulmonary syndrome, Chronic sinusitis, Plague are the diseases that belongs to most communities in the network.

We have analysed clusters relatedness using linkcomm libraray in R, 83 overlappings clusters found as shown in dendogram figure 5. Furthermore, the symptoms i.e. Abdominal pain, Diarrhea, Fatigue, Headache, and Nausea and vomiting are the most common symptoms which depicts the similarity between diseases in overlapping communities detection.

5. Conclusion and Future Work

We have examined disease comorbidity patterns based on their symptoms using network analysis technique. Mostly, diseases are connected because of their shared symptoms. Our analysis, revealed that the diseases Ascariasis, Churg-Strauss syndrome, HIV/AIDS, Aspergillosis, and Wegener's granulomatosis are the most co-occurred diseases in the network. The future work of this study includes the disease association patterns on longitudinal data to find disease comorbidity and evolution patterns.

Fig. 3 - Community Detection

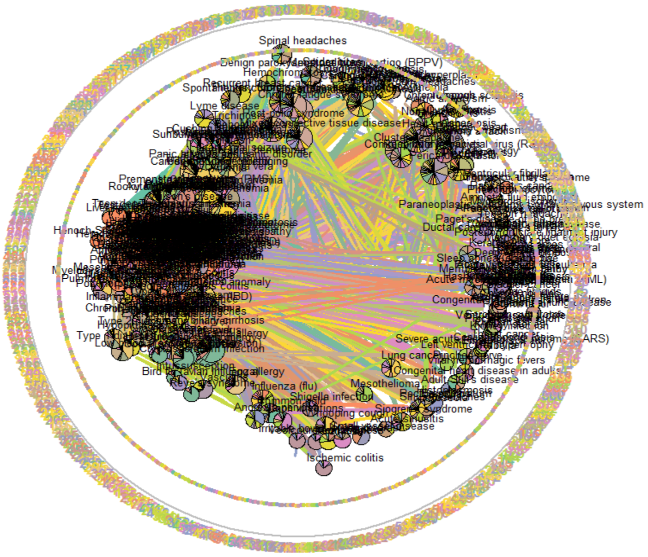
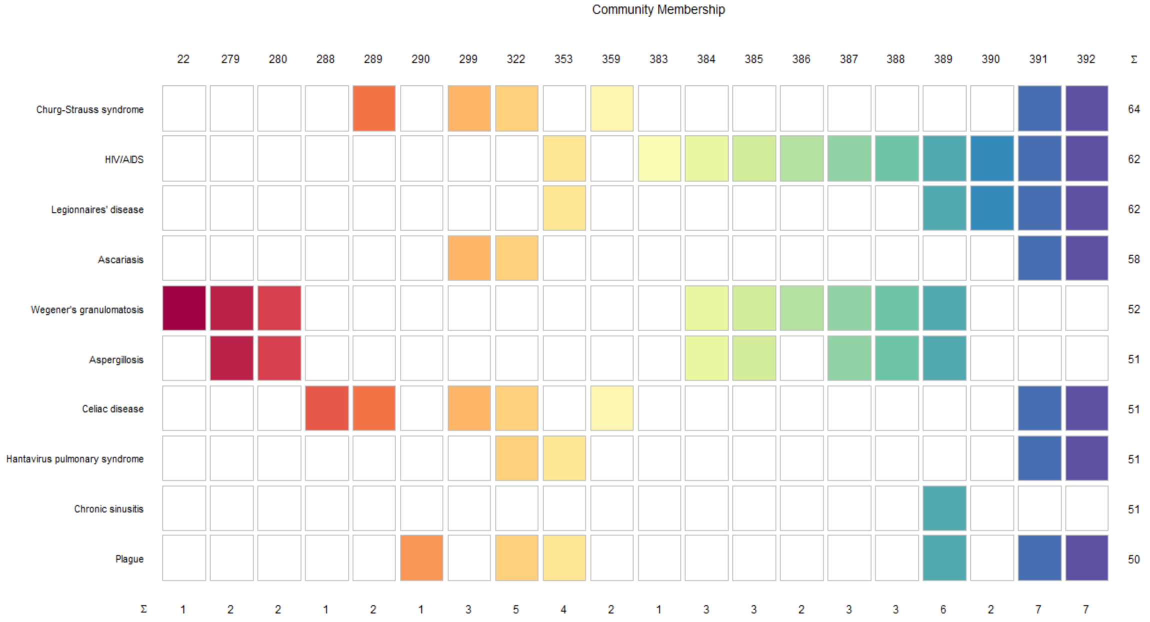
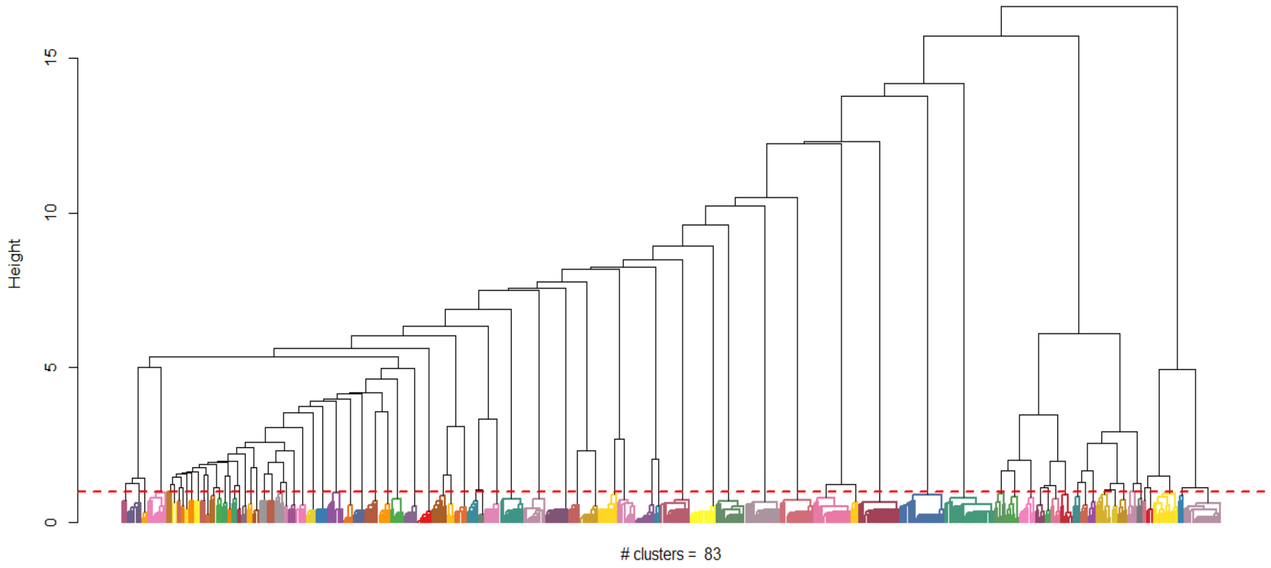


Fig. 4 - Community Membership Matrix



Received: 08.06.2021 – Accepted: 09.07.2021

Fig. 5 - Dendogram



References

X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, “ARTICLE Human symptoms–disease network,” *Nat. Commun.*, vol. 5, 2014.

K. Sun, J.P. Gonçalves, C. Larminie, and N. Przulj, “Predicting disease associations via biological network analysis,” *BMC Bioinformatics*, vol. 15, p. 304, 2014.

J.H. Kim, K.Y. Son, D.W. Shin, S.H. Kim, J.W. Yun, J.H. Shin, M.S. Kang, E.H. Chung, K.H. Yoo, and J.M. Yun, “Network analysis of human diseases using Korean nationwide claims data,” *J. Biomed. Inform.*, 61, 276–282, 2016.

K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, and A.L. Barabási, “The human disease network,” *Proc. Natl. Acad. Sci. U.S.A.*, 104(21), 8685–8690, 2007.

K. Il Goh and I.G. Choi, “Exploring the human diseasome: The human disease network,” *Brief. Funct. Genomics*, 11(6), 533–542, 2012.

O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, “Associating genes and protein complexes with disease via network propagation,” *PLoS Comput. Biol.*, 6(1), 2010.

D.S. Lee, J. Park, K.A. Kay, N.A. Christakis, Z.N. Oltvai, and A.L. Barabasi, “The implications of human metabolic network topology for disease comorbidity,” *Proc Natl Acad Sci USA*, 105(29), 9880–9885, 2008.

X. Wu, Q. Liu, and R. Jiang, “Align human interactome with phenome to identify causative genes and networks underlying disease families,” *Bioinformatics*, vol. 25, no. 1, pp. 98–104, 2009.

M. Ghadie and Y. Xia, “Estimating dispensable content in the human interactome,” *Nat. Commun.*, 10(1), 3205, 2019.

Y. Li and P. Agarwal, “A pathway-based view of human diseases and disease relationships,” *PLoS One*, 4(2), 2–7, 2009.

F.S. Roque *et al.*, "Using electronic patient records to discover disease correlations and stratify patient cohorts," *PLoS Comput. Biol.*, vol. 7, no. 8, 2011.

D. Lee, M. Kim, and H. Shin, "Inference on chains of disease progression based on disease networks," *PLoS One*, vol. 14, no. 6, p. e0218871, 2019.

Z. Batool, M. Usman, K. Saleem, M. Abdullah-Al-Wadud, Fazal-e-Amin, and A. Al-Eliwi, "Disease–Disease Association Using Network Modeling: Challenges and Opportunities," *J. Med. Imaging Heal. Informatics*, vol. 8, no. 4, pp. 627–638, 2018.

J. Menche *et al.*, "Disease networks. Uncovering disease-disease relationships through the incomplete interactome.," *Science (80-.).*, 347(6224), 1257601, 2015.

P. Björntorp, "'Portal' adipose tissue as a generator of risk factors for cardiovascular disease and diabetes.," *Arterioscler. An Off. J. Am. Hear. Assoc. Inc.*, 10(4), 493–496, 1990.

Received: 08.06.2021 – Accepted: 09.07.2021