Topological Inquisition into the PPI Networks Associated with Human Diseases Through Graphlet Frequency Distribution

Debjani Bhattacharjee[1], Sk Md Mosaddek Hossain[2](✉), Raziya Sultana[2], and Sumanta Ray[2](✉)

[1] Acharya Prafulla Chandra College, New Barrackpore,
Kolkata 700131, West Bengal, India
[2] Department of Computer Science and Engineering, Aliah University,
New Town, Kolkata 700156, West Bengal, India
mosaddek.hossain@gmail.com, sumanta.ray@aliah.ac.in

Abstract. In this article, we have proposed a new framework to compare topological structure of protein-protein interaction (PPI) networks constructed from disease associated proteins. Here, similarity of local topological structure between networks is discovered through the analysis of frequent sub-pattern occurred in them using a novel similarity measure based on graphlet frequency distribution. Graphlets are small connected non-isomorphic induced subgraphs in a network which provides detailed topological statistics of it. We have analyzed pairwise similarity of 22 disease associated PPI networks and compared topological and biological characteristics. It has been observed that the PPI networks associated with disease classes 'metabolic' and 'neurological' have the highest similarity scores. Higher similarity has also been observed for networks of disease classes 'bone' and 'skeletal'; 'endocrine' and 'multiple'; and 'gastrointestinal and respiratory'. Topological analysis of the networks also reveals that degree and betweenness centrality of proteins is strongly correlated for the network pairs with high similarity scores. We have also performed gene ontology and pathway based analysis of the proteins involved in the disease associated networks.

1 Introduction

Analyzing and understanding the intricate structure of Human Disease Network (HDN) is one of the most challenging fields in computational biology research [1]. Most human diseases are complex as they are not only associated with a single gene but a group of genes [2]. A comprehensive study on disease similarities provides new ideas about the cause of diseases and act as the key player in diagnosis and treatment of these complex diseases [3]. In [4], a disease phenotype network is constructed by performing a text mining approach to group common clinical terms. Goh et al. [1], first introduce the concept of human-disease network which provides a network of disorders and disease genes which are linked with known

B.U. Shankar et al. (Eds.): PReMI 2017, LNCS 10597, pp. 431–437, 2017.
https://doi.org/10.1007/978-3-319-69900-4_55

gene disease association. Using this data Bandyopadhyay et al. [5] proposed an approach to find disease associated protein complexes in human PPI network. In [6], a novel framework is introduced to discover the similarity between two tissue or disease specific networks through multi-label graphlet counting. In [7], novel disease gene association is predicted by using RWR algorithm and functional similarity between protein complexes. A novel framework is proposed to compare biological networks using graphlet degree distribution in [8].

Here, we proposed a novel framework to compare the local structure of disease associated PPI networks using graphlet frequency. For this, first we have constructed 22 PPI networks from the 22 disease/disorder classes reported in [1]. Each PPI network consists of proteins associated with a particular disorder/disease class. We have found the occurrences of 3-, 4-, and 5-node graphlets in each of the networks and compare the occurrences to know the topological similarity between two networks. To count the occurrences of graphlets we have utilized a widely popular tool called G-trie Scanner [9]. In G-trie scanner a tree is constructed with set of sub graphs based on common structure or patterns in which nodes are connected. We have proposed a similarity measure which take the occurrences of graphlets and return a similarity score which signifies the similarity between the network structure. We have analyzed the similarity scores of each pair of networks, predict disease pairs having high similarity between the associated network structure. We have also analyzed topological properties of each network and conducted a gene ontology and pathway based analysis.

2 Method

This section describes the proposed framework to compare topological similarity between disease associated PPI Networks.

2.1 Dataset Preparation

We have downloaded the disease gene association database from Goh et al. [1]. The dataset is modeled by a bipartite network consisting of two disjoint sets of nodes: one set represents disease/disorder whereas the other sets corresponds to associated genes. The disorders/disease list and the responsible genes are collected from Online Mendelian Inheritance in Man (OMIM; [10]), a repository of human disease genes and phenotypes. In [1], all the disease/disorders are categorized into 22 broad classes. We have utilized this data and mapped all the disease associated proteins in human PPI network downloaded from Human Protein Reference Database (HPRD) [11]. Thus we get 22 PPI networks, each of which consists of proteins associated with a particular disease class. All the networks are highly sparsed and the density ranging from 6.9067e−06 to 1.0116e−04. We have utilized DAVID Functional Annotation Bioinformatics tool [12] for functional enrichment analysis.

2.2 Comparing the PPI Networks Using Graphlet Frequency

For each network, we have computed the occurrences of 29 graphlets using G-trie scanner [9]. Here, we have considered the graphlets of node size 3, 4 and 5 for comparison purpose as shown in Fig. 1-(a). We have computed similarities between two disease associated networks by comparing the occurrences of graphlets. Since the network size for each disease category is different, we normalized the occurrences by dividing each occurrence by its respective network size. Next, we arranged all the obtained graphlet frequencies in a 22 × 29 adjacency matrix, where we have 22 categories of diseases and frequency of 29 graphlets structure for each disease network. Let k be the number of graphlets (here, $k = 29$), $N_1(G_i)$ represents the occurrence of graphlet G_i in network N_1 and $N_2(G_i)$ represents the same for network N_2, then we have computed similarity score between two networks as:

$$sim(N_1, N_2) = \frac{\sum_{i=1}^{k} \frac{\min(N_1(G_i), N_2(G_i))}{\max(N_1(G_i), N_2(G_i))}}{k}. \tag{1}$$

The similarity scores *sim* is equal to 0 for two exactly same networks and maximum 1 for two networks having maximum disagreement in terms of graphlets occurrences. Network similarity is finally estimated by comparing the similarity score between the networks using the equation above. Thus we get a similarity matrix of dimension 22 × 22 which represents the pairwise similarity between two disease associated networks.

3 Results

3.1 Comparing Networks Using Similarity Score

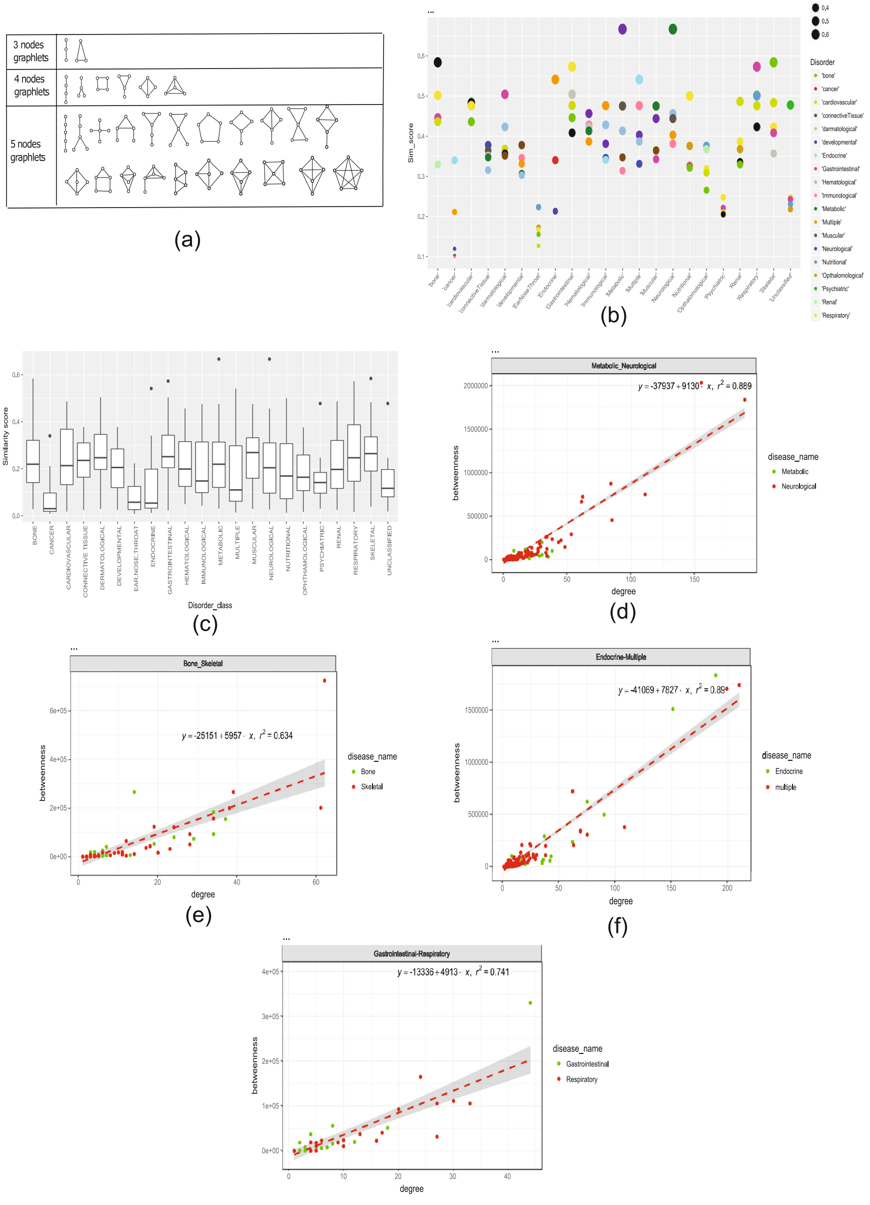
We have compared the topological structure of the disease associated networks using the similarity score specified in Eq. (1). First, 22 PPI networks are formed from each disease associated protein set. Next, for each pair of networks, similarity score is identified. High score between two networks signifies that the 29 graphlets follows same patterns within the two networks. Similarly, low score represents that there is an inconsistency between the occurrences of the graphlets within two networks. We have computed the similarity scores between each pair of networks and depicted these in Fig. 1-(b) and (c) with a box plot. From this figure, it can be observed that the network associated with 'mascular' and 'cancer' disease classes have high and low similarity scores with other networks, respectively. This suggests that the topological structure of PPI network associated with cancer disease class is dissimilar to other disease associated networks. In Fig. 1-(b), we have shown a visualization of the similarity structure between networks using the similarity scores. For each network associated with a disease class, we have chosen top five networks having high similarity value and plotted these. Here, color and size of each circle is varying with disease class and similarity scores, respectively. From the Fig. 1-(b), it is observed that network structure of disease classes 'metabolic' and 'developmental' has high similarity value. Similarly the following network pairs have high similarity scores: 'bone-skeletal', 'endocrine-multiple', 'gastrointestinal-respiratory' and 'metabolic-neurological'.

3.2 Topological Analysis of Disorder Associated Proteins

To investigate whether the similarity scores are correlated with the topological features of the proteins involved within the networks, we find degree and betweenness centrality of each protein associated with the 22 disease associated networks. Here, degree of a protein signifies number of interactions it made within the whole human interactome. Betweenness centrality of a protein is also calculated by considering the whole human PPI network. We observe that degree and betweenness centrality is strongly correlated for the disease associated PPI networks with high similarity scores. Figure 1-(d–g), shows the scatter plots between degree and betweenness centrality of proteins associated with two disease classes. We have taken four pairs of disease associated networks with high similarity scores and plot degree vs. betweenness centrality of each nodes. It can be seen from the figure that disease pair 'metabolic-neurological' having the highest similarity score 0.67, has a strong correlation between degree and betweenness centrality of associated proteins($R^2 = 0.89$). The similar results can be observed for disease pairs: 'bone-skeletal' ($sim\_score = 0.583$, R^2 value 0.634), 'gastrointestinal-respiratory' ($sim\_score = 0.573$, R^2 value 0.741), and 'endocrine-multiple' ($sim\_score = 0.541$, R^2 value 0.89). To know whether there is any difference in degree or betweenness centrality of proteins associated with the similar disease pairs we plot these two metric for each protein which are associated with four similar disease pairs: 'bone-skeletal', 'endocrine-multiple', 'gastrointestinal-respiratory' and 'metabolic-neurological'. Figure 1-(h–i) show the box and jitter-plot of degree and betweenness centrality of those proteins. It can be observed from the figures that there is no distinguishable difference in degree and betweenness centrality of proteins associated with the similar disease pairs.

3.3 Functional Enrichment Analysis

Gene ontology based analysis is the most important and strong tool to identify the underlying biological meaning and functions of a set of proteins. Here, we have investigated the gene ontology terms and pathways which are associated with the proteins of the most similar disease associated network pairs: 'bone-skeletal', 'endocrine-multiple', 'gastrointestinal-respiratory' and 'metabolic-neurological'. In Table 1, the most significant GO-terms and KEGG pathway are listed. As can be seen from the table that similar disease pair like 'gastrointestinal-respiratory' is enriched in same biological process 'MAPK cascade'. Most of the disease associated proteins are enriched in cancer pathways.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 1. Panel (a) shows the structure of 29 graphlets. Panel (b) and panel (c) represents dot plot and box plot of similarity scores of all the disorder classes with other classes, respectively. Color and size of the dots are varying with respect to the disorder class and similarity scores shown in the legends. Panel (d–g) represents scatter plots of correlations between degree and betweenness centrality of disease associated proteins for the disease pairs: 'bone-skeletal', 'endocrine-multiple', 'gastrointestinal-respiratory' and 'metabolic-neurological'. Panel-(h–i) represents Box and Jitter plots of the same for the four disease pairs.

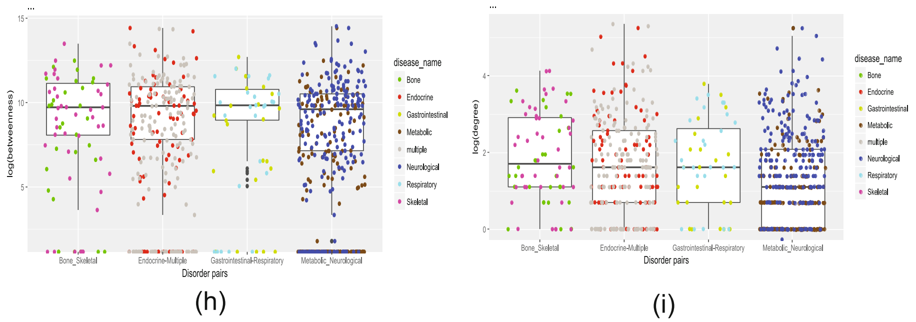


Fig. 1. (*continued*)

Table 1. Table shows the gene ontology terms and KEGG pathway associated with the proteins of most similar pair of disease classes: ‘bone-skeletal’, ‘endocrine-multiple’, ‘gastrointestinal-respiratory’ and ‘metabolic-neurological’

Disease_class	GO-term_(GO-id)/p-value	KEGG_Pathway/ p-value
Bone	Extracellular matrix organization (GO:0030198)/1.60E−32	Pathways in cancer (6.50E−18)
Skeletal	Positive regulation of transcription from RNA polymerase II promoter (GO:0045944)/1.40E−36	Pathways in cancer (3.90E−32)
Endocrine	Positive regulation, of transcription from RNA polymerase II promoter (GO:0045944)/1.90E−69	Pathways in cancer (1.60E−32)
Metabolic	Positive regulation, of transcription from RNA polymerase II promoter (GO:0045944)/1.70E−27	Prostate cancer (1.80E−16)
Gastrointestinal	MAPK cascade (GO:0000165)/2.40E−14	ErbB, signaling pathway (1.80E−14)
Respiratory	MAPK cascade (GO:0000165)/8.40E−20	Proteoglycans in cancer (1.50E−20)
Neurological	Not found	Not found

4 Conclusions

In this paper, we have proposed a novel framework to compare the topological structure of disease associated PPI networks. It appears from the analysis that the PPI networks corresponding to the disease pair ‘bone-skeletal’, ‘endocrine-multiple’, ‘gastrointestinal-respiratory’ and ‘metabolic-neurological’ are similar with respect to their topological features. It is also observed from the topological analysis of the disease associated proteins that degree and betweenness centrality is strongly correlated for similar disease associated network pair. Functional enrichment analysis also reveals that the proteins associated with similar disease associated networks pair are enriched in same gene ontology terms. Further analysis and a proper investigation of biological properties of similar and dissimilar disease associated PPI networks may yield some new insights into the underlying structure of disease-gene association.

References

1. Goh, K., Cusick, M., Valle, D., Childs, B., Vidal, M., Barabasi, A.: The human disease network. PNAS **104**, 8685–8690 (2007)
2. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M.: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. **38**(1), D355 (2010)
3. Yang, J., Wu, S.-J., Yang, S.-Y., Peng, J.-W., Wang, S.-N., Wang, F.-Y., Song, Y.-X., Qi, T., Li, Y.-X., Li, Y.-Y.: DNetDB: the human disease network database based on dysfunctional regulation mechanism. BMC Syst. Biol. **10**(1), 36 (2016)
4. Chen, Y., Zhang, X., Zhang, G.-Q., Xu, R.: Comparative analysis of a novel disease phenotype network based on clinical manifestations. J. Biomed. Inf. **53**, 113–120 (2015)
5. Bandyopadhyay, S., Ray, S., Mukhopadhyay, A., Maulik, U.: A multiobjective approach for identifying protein complexes and studying their association in multiple disorders. Algorithms Mol. Biol. **10**(24) (2014). doi:10.1186/s13015-015-0056-2
6. Sonmez, A.B., Can, T.: Comparison of tissue/disease specific integrated networks using directed graphlet signatures. In: Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, ser. BCB 2016, New York, NY, USA, pp. 533–534. ACM (2016)
7. Le, D.-H.: A novel method for identifying disease associated protein complexes based on functional similarity protein complex networks. Algorithms Mol. Biol. **10**(1), 14 (2015). https://doi.org/10.1186/s13015-015-0044-6
8. Pržulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics **23**(2), e177 (2007)
9. Ribeiro, P., Silva, F.: G-tries: a data structure for storing and finding subgraphs. Data Mining Knowl. Discov. **28**(2), 337–377 (2014)
10. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. **33**(Database issue), 514–517 (2005)
11. Prasad, T., Goel, R., Kandasamy, K.: Human protein reference database. Nucleic Acids Res. **37**, D767–D772 (2009)
12. Huang, D., Sherman, B., Tan, Q., Collins, J., Alvord, W., Roayaei, J., Stephens, R., Baseler, M., Lane, H., Lempicki, R.: The David gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. **8**(9), R183 (2007)