Is the Average Shortest Path Length of Gene Set a Reflection of their Biological Relatedness?

Varsha Embar\(^1\), Adam Handen\(^2\), and Madhavi K. Ganapathiraju\(^{1,2,*}\)
\(^1\)Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 15213 USA
\(^2\)Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, 15213, USA

Abstract

When a set of genes are identified to be related to a disease, say through gene expression analysis, it is common to examine the average distance among their protein products in the human interactome as a measure of biological relatedness of these genes. The reasoning for this is that, genes associated with a disease would tend to be functionally related, and that functionally related genes would be closely connected to each other in the interactome. Typically, average shortest path length (ASPL) of disease genes is compared to ASPL of randomly selected genes or to ASPL in a randomly permuted network. We examined whether the ASPL of a set of genes is indeed a good measure of biological relatedness or whether it is simply a characteristic of the degree distribution of those genes. We examined the ASPL of genes sets of some disease and pathway associations and compared them to ASPL of three types of randomly selected control sets: uniform selection, from entire proteome, degree-matched selection and random permutation of the network. We found that disease associated genes and their degree-matched random genes have comparable ASPL. In other words, ASPL is a characteristic of the degree of the genes and the network topology, and not that of functional coherence.

1. Background

Protein-protein interaction (PPI) networks are undirected graphs that denote physical interaction between proteins. They are scale free networks i.e., the number of interactions of a protein follows the power law distribution which has been a noticeable phenomenon over the years as shown in Figure 1 [1]. Graph theory has been applied to PPI networks to analyze and predict biological phenomena using topological features such as clustering coefficients, presence of highly connected hubs, modular organization, etc [2, 3].

Average shortest path length (ASPL), a topological feature of scale-free networks, is a commonly used metric for the analysis of disease-associated genes and is used to demonstrate whether the genes are functionally cohesive. The intuition behind this is that the

\* Although referred to as ‘genes’ in the context of disease associations, it is the protein-products that are represented in the interactome. Thus, throughout the paper, ‘ASPL among genes’ refers to ‘ASPL among their protein products’.

\*Corresponding Author: madhavi@pitt.edu.

Conflict of Interest

Authors declare that there is no conflict of interest
genes that are functionally related are going to be closer to each other in the interactome. Typically, a low ASPL is reported between candidate genes compared to randomly selected genes along with an empirical p-value of statistical significance [4–9]. ASPL is commonly used in gene-prioritization [4, 10]. It has been used to study pathways in complex diseases to connect putative causal genes to target genes and to extract disease-associated network modules [11].

However, it has been shown mathematically that in scale-free networks, the ASPL of the network converges to a value that is solely dependent on the number of nodes and the average degree of nodes [12]. This property has been empirically shown to hold true for PPI network of yeast. Comparing the ASPL of yeast PPI network to ASPL of randomly generated degree-matched networks showed that they converged to the same value [13].

This leads us to this question: is the ASPL of a subset of genes a reflection of their biological relatedness or that of the degree of those genes? In this work, we examined the ASPL of typical candidate genes sets, namely those that are associated with diseases, with randomly selected genes and in randomly permuted network. We also perform experiments to study the effect of hub proteins in characterizing the ASPL of the candidate gene sets.

2. Methods

We collected human protein-protein interactions from two databases: BioGRID and HPRD [14, 15]. We filtered the data set to only include direct interactions (MI:0407 in the PSI MI Ontology). This yielded 77,534 non-redundant interactions among 13,217 genes.

Pathways were acquired from REACTOME and diseases from KEGG [16, 17]. We filtered the annotated gene sets to include only those with a minimum of 10 genes. This yielded 1,040 pathways and 283 diseases among 7,305 and 5,396 genes respectively. We also analyzed six traits studied in Genome Wide Association Studies (GWAS): Type 2 diabetes, celiac disease, bipolar disorder, obesity, Alzheimer’s disease and Crohn’s disease. We refer to the genes associated with a given annotation as ‘candidate genes’.

Average length of shortest path between every pair of candidate genes is computed. In the event that there is no path between two genes, we defined that shortest path length to be 5 more than the largest distance observed between any pair of genes in the network, thus setting it 17. We used NetworkX, a Python library, for computing ASPL [18].

We compared the ASPL of candidate genes in the human interactome with that in a randomly permuted network. We generated the permuted network by swapping the edges in the graph while still keeping the degree of each node fixed.

We also compared the ASPL of candidate genes with that of randomly selected genes in the human interactome. We employed two methods to sample genes randomly:

- Uniform sampling: Uniform random sampling from the human genes from a total 19,676 genes in HPRD database.
• Degree-matched random sampling: Random genes were sampled with the constraint that they have (nearly) the same degree distribution as the candidate genes. We binned all the genes according to their degree in the interactome using varied bin sizes (Table 1), and then sampled the same number of genes from each bin as the number of candidate genes in that bin, while ensuring that candidate genes are not included in the sample.

3. Results

Table 2 provides the general architecture of the Human PPI network with overall statistics about its size, density and connectivity. The interactome has 30 connected components. There are only 13,217 human genes with known interactions, leaving thousands of genes without any known interactions. The ASPL values for each of the sampling methods is reported in Table 3.

When compared with random genes selected with uniform sampling, the disease or pathway associated genes always had a significantly lower ASPL, suggesting that they were much more closely connected in the network. However, when compared with random genes that were constrained to have same degree as candidate genes, there was no discernible difference in their ASPL. This was also true when the comparison was made to candidate genes in a randomly permuted network. Thus, it is seen consistently that, any gene sets that had same degree distribution in a given network or another network with same global degree distribution, had similar ASPLs.

For each of the three methods, we computed the ASPL for 1000 gene sets for each annotation. We performed a two sided t-test which showed that all of the traits had significant p-values when compared against ASPL of random gene lists and found that the results had a high statistical significance.

Next, we considered the possibility that the presence of hub proteins may influence the ASPL values of the selected gene sets. In order to account for it, we carried out similar analysis as before, but by removing hub genes from the gene sets. First we considered removing hub genes with degree >50, and then hub genes with degree >25. Despite removing these hub genes from the candidate and random gene sets, it remains that the degree-matched random genes have comparable ASPL as candidate genes (Figure 2).

4. Conclusions

ASPL is a commonly used metric to show that a candidate set of genes is functionally related in many system biology works. In this paper, we discuss the bias that the method may hold due to the incomplete nature of PPI networks and also because some genes may be better studied than others due to their association with diseases. We performed an experiment by using three different sampling methods and compared the ASPL of the sample genes with a candidate set of genes known to be associated with diseases and pathways. We found that the ASPL of the degree matched random genes and the ASPL of genes in the permuted network converge to the same value as that of the candidate genes. We examined whether the presence of hub genes in a candidate gene set, and therefore a
corresponding gene in the degree-matched random set, was the reason for comparable ASPL values, but concluded the ASPL value of degree-matched random genes remained to be similar to that of candidate genes even after removing hub genes.

We conclude that the AS is more a characteristic of the network topology rather than that of biological cohesiveness. That is, even a randomly selected set of genes can be as closely connected as biologically related genes (ex. genes involved in a disease), as long as the random genes have same degree distribution as the candidate genes.

Acknowledgments

This work is funded by the BRAINS grant R01MH094564 from the National Institute of Mental Health of the National Institutes of Health (NIMH/NIH), USA.

References

1. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. May 3; 2001 411(6833):41–2. [PubMed: 11333967]
2. Krawetz, SA., Werner, T. Bioinformatics for systems biology. Springer; 2009.
3. Joy MP, Brock A, Inger DE, Huang S. High-betweenness proteins in the yeast protein interaction network. BioMed research international. 2005; 2005(2):96–103.
4. Zhang L, Li X, Tai J, Li W, Chen L. Predicting candidate genes based on combined network topological features: a case study in coronary artery disease. PloS one. 2012; 7(6):e39542. [PubMed: 22761820]
5. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. The American Journal of Human Genetics. 2008; 82(4):949–958. [PubMed: 18371930]
6. Franke L, Van Bakel H, Fokkens L, De Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. The American Journal of Human Genetics. 2006; 78(6):1011–1025. [PubMed: 16685651]
7. Choura M, Rebai A. Topological features of cancer proteins in the human NR-RTK interaction network. Journal of Receptors and Signal Transduction. 2012; 32(5):257–262. [PubMed: 22765391]
8. Isik Z, Baldow C, Cannistraci CV, Schroeder M. Drug target prioritization by perturbed gene expression and network information. Sci Rep. 2015; 5:17417. [PubMed: 26615774]
9. Ni Q, Su X, Chen J, Tian W. Prediction of Metabolic Gene Biomarkers for Neurodegenerative Disease by an Integrated Network-Based Approach. BioMed research international. 2015
10. Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, Mooney SD. An integrated approach to inferring gene-disease associations in humans. Proteins. Aug 15; 2008 72(3):1030–7. [PubMed: 18300252]
11. Cho DY, Kim YA, Przytycka TM. Chapter 5: Network biology approach to complex diseases. PLoS Comput Biol. 2012; 8(12):e1002820. [PubMed: 23300411]
12. Chung, FR., Lu, L. Complex graphs and networks. American mathematical society; Providence: 2006.
13. Gibson, TA., Goldberg, DS. Algorithms in Bioinformatics. Springer; 2015.
14. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic acids research. 2006; 34(suppl 1):D535–D539. [PubMed: 16381927]
15. Wilson N. Human Protein Reference Database. Nature Reviews Molecular Cell Biology. 2004; 5(1):4–4.
Biographies

Varsha Embar received B.S. in Computer Science and Engineering from M. S. Ramaiah Institute of Technology, India. She is currently a graduate student and a research assistant at the Language Technologies Institute at Carnegie Mellon University, Pittsburgh, USA pursuing an M.S in Language Technologies.

Adam Handen received B.S. and M.S. degrees in Bioinformatics from Rochester Institute of Technology, USA in 2014. He received M.S. degree in Biomedical Informatics from University of Pittsburgh in 2015.

Madhavi K. Ganapathiraju received B.Sc. degree from Delhi University, India, M.Engg degree from Department of Electrical Communications Engineering, Indian Institute of Science, and Ph.D. in Language and Information Technologies from School of Computer Science, Carnegie Mellon University in 2007. She is currently Associate Professor at Department of Biomedical Informatics with Secondary Appointment in Intelligent Systems Program, University of Pittsburgh, USA. She holds affiliated faculty position at Language Technologies Institute, Carnegie Mellon University. She is Member of IEEE.
Figure 1.
Degree distribution of genes in the human interactome: Most of genes with high degree are associated with diseases imply a literature bias towards the genes studies.
Figure 2.
ASPL after removing hub genes (a) without genes with degree greater than 50 (b) without genes with degree greater than 25
Table 1

The range of degrees a node may have: There are only 13,217 human genes with known interactions from HPRD and BioGRID, leaving most of the genome without any known interactions.

| Bin Size | Genes |
|----------|-------|
|          |       |
| 0        | 34500 |
| 1        | 2612  |
| 2–3      | 3043  |
| 4–10     | 3946  |
| 11–25    | 2233  |
| 26–50    | 858   |
| 51–75    | 257   |
| >75      | 268   |
**Table 2**
Topological properties of the Human interactome

| Property                      | Value |
|-------------------------------|-------|
| Number of nodes               | 13,217|
| Number of edges               | 77,534|
| Clustering co-efficient       | 0.09  |
| Average degree connectivity   | 59.3  |
| Number of connected components| 30    |
| Average number of neighbors   | 11.7  |
| Density                       | 0.0009|
| Diameter                      | 12    |
Table 3

Comparison of shortest path distances: Average shortest path lengths are shown for different genes sets in the computed in the human interactomes network (columns C–E) and in randomly permuted network (column F). Distances have been rounded to closest integer value so that it is easy to see the similarities.

| Annotation                              | #Genes | Candidate Genes | Random Genes | Degree Matched Random | Permuted Network |
|-----------------------------------------|--------|-----------------|--------------|-----------------------|-----------------|
|                                         |        | A               | B            | C                     | D               | E               | F               |
| **KEGG Diseases**                       |        |                 |              |                       |                 |                 |
| Type 1 diabetes mellitus                | 22     | 6               | 11           | 6                     | 6               |
| Dilated cardiomyopathy                  | 16     | 3               | 11           | 3                     | 3               |
| Deafness, autosomal recessive           | 24     | 12              | 11           | 12                    | 12              |
| Hereditary spastic paraplegia           | 19     | 11              | 11           | 11                    | 11              |
| Gastric cancer                          | 20     | 3               | 11           | 3                     | 3               |
| **Reactome Pathways**                   |        |                 |              |                       |                 |
| Olfactory Signaling Pathway             | 374    | 17              | 11           | 17                    | 17              |
| Hexose transport                        | 43     | 8               | 11           | 8                     | 8               |
| Eukaryotic Translation Elongation       | 89     | 5               | 11           | 5                     | 5               |
| Mitochondrial biogenesis                | 53     | 3               | 11           | 4                     | 3               |
| Global Genomic NER                     | 35     | 3               | 11           | 4                     | 4               |
| Nuclear Envelope Breakdown              | 44     | 6               | 11           | 6                     | 6               |
| Toll Like Receptor 3 Cascade           | 98     | 3               | 11           | 3                     | 3               |
| Viral mRNA Translation                  | 83     | 4               | 11           | 5                     | 5               |
| FRS2-mediated cascade                   | 35     | 4               | 11           | 5                     | 5               |
| **GWAS Diseases**                       |        |                 |              |                       |                 |
| Type 2 diabetes                         | 103    | 8               | 11           | 8                     | 8               |
| Celiac disease                          | 64     | 7               | 11           | 7                     | 7               |
| Schizophrenia (GWAS) [19]               | 76     | 11              | 11           | 11                    | 11              |
| Bipolar disorder                        | 128    | 10              | 11           | 10                    | 10              |
| Crohn’s disease                         | 207    | 9               | 11           | 9                     | 8               |
| Obesity                                 | 51     | 8               | 11           | 8                     | 8               |
| Alzheimer’s disease                     | 36     | 8               | 11           | 8                     | 8               |
| Annotation                                      | #Genes | Candidate Genes | Random Genes | Degree Matched Random | Permutated Network |
|------------------------------------------------|--------|-----------------|--------------|-----------------------|--------------------|
| Acute lymphoblastic leukemia (childhood)       | 24     | 7               | 11           | 7                     | 7                  |
| OMIM Diseases                                   |        |                 |              |                       |                    |
| Leber congenital amaurosis                      | 15     | 9               | 11           | 9                     | 9                  |
| Bardet-Biedl syndrome                           | 14     | 7               | 11           | 7                     | 7                  |
| Fanconi anemia                                  | 14     | 2               | 11           | 3                     | 3                  |
| Limb-girdle muscular dystrophy                  | 17     | 10              | 11           | 8                     | 8                  |
| Retinitis pigmentiosa                           | 45     | 10              | 11           | 9                     | 9                  |
| Other                                           |        |                 |              |                       |                    |
| Schizophrenia (Historical) [19]                 | 25     | 5               | 11           | 5                     | 5                  |