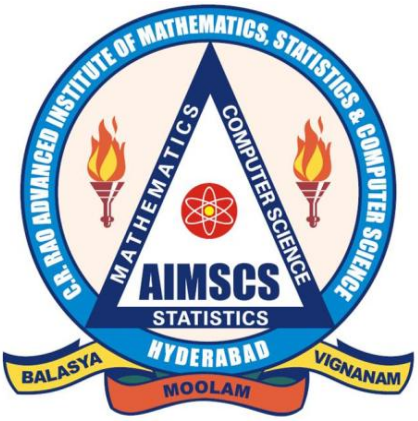
CRRAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS)

Research Report



Author (s): **T Kranthi, S B Rao, and P Manimaran**

Title of the Report: **Prediction of essential genes in Human using graph centrality measures and a machine learning approach**

Research Report No.: **RR2014-28**

Date: **November 20, 2014**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

Prediction of essential genes in Human using graph centrality measures and a machine learning approach

T Kranthi, S B Rao, and P Manimaran

C R Rao Advanced Institute of Mathematics, Statistics and Computer Science, University of Hyderabad Campus, Prof. C R Rao Road, Gachibowli, Hyderabad – 500046, India.

Abstract

Recent technological advances in experimental biology have yielded massive amounts of biological data which made its analysis a highly onerous task. Networks are inestimable models for ameliorated analysis and efficient interpretation of biological systems and the mathematical discipline which underpins the study of these complex biological networks is graph theory. In our present work we have tried to accentuate the importance of applications of graph theory on biological systems through the prediction of Human essential genes utilizing the combined centrality and machine learning approach. Of the predicted essential genes 854 genes were found to be in common with 1950 known essential genes. The essentiality of the remaining genes was also corroborated through the literature survey and thus our work directs the attention of application of graph theoretical approaches on biological systems.

Keywords: Graph theory, Protein interactions, Centrality measures, Machine learning, Human essential genes.

1. Introduction

Majority of the systems either available in nature or manmade are complex. Understanding these complex systems requires a bottom up approach i.e. breaking the system into small elementary constituents. Mapping out the interactions between these components can be characterized as network. Historically, the study of networks has been mainly the domain of a branch of discrete mathematics known as graph theoryhas become the fundamental pillars of discrete mathematics. In view of graph theory, network can be defined as any real system natural or artificial that can be completely described by a mathematical graph, an object composed of vertices connected by edges which can be mathematically represented as a graph composed of vertices 'V' and edges 'E' is G(V,E). Many different network models have been proposed to address properties of complex systems, some of the important ones being the random network model proposed by Erdos and Renyi, the small world network proposed by Watts and Strogatz, and the scale-free model proposed by Barabasi and Albert. The failure in studying the real world networks further paved the way for Watts and Strogatz to propose the 'small world network', a network with a small diameter and high clustering. More recently, Barabasi and Albert have proposed the scale free model, in which the degree distribution which is nothing but the property of resilience of networks to the removal of their vertices possesses power law. The networks can also be classified as directed and undirected networks. A network is directed if all of its edges are directed and undirected network can be represented by a directed one having two edges between each pair of connected vertices, one in each direction. The networks can also be further classified as weighted and unweighted networks depending on the weights assigned to their edgesUtilizing graph theory concepts networks enable a simple and uniform representation of complex structures, processes and finds wide range of applications in various fields such as social, physical and biological sciences [1-8].

Mathematically, a network or a graph can be represented in the form of a matrix called Adjacency matrix (A) in which a_{ij}=1 if 'i' and 'j' are connected with an edge. For undirected network, the adjacency matrix is symmetric square matrix whereas for directed network it is asymmetric square matrix. For an unweighted network, the weights are unknown and are represented as "unit weight" i.e. 1 to all the edges of the network. It just represents the connection between two nodes. In the weighted network, the weights are given to the edges

between a pairs of nodes. The weights in numeric for the edges may represent distance, transmission speed, reaction rate, interaction time etc.The topological properties such as degree distribution, clustering coefficients, centrality measures, community structures, modularity etc. of the network models help us to understand the overall properties through their structure [9-12]. The characteristicsof individual nodes are essential during the analysis of the importance of nodes in terms of connectivity, information transfer capability and closeness toother nodes etc. In order to establish the individual properties of a node centrality measures have been proposed. The graph centrality measure concept plays a vital role in identifying the potential nodes that are functionally important in a network. In the recent past, various centrality measures such as degree, closeness, betweenness, Eigen vector, information centrality etc. have been developed for predicting the potentiality of a node [13-18].

All the biological networks are scale-free in nature, which suggest that only a small number of nodes are highly connected, whereas a large number of nodes have fewer connections. Consequently, onlythe small number of nodes that have many connections,referred to as ‘hubs’, control the overall robustness of thenetwork. The above mentioned centrality measures aids us in analyzing the various underlying process and robustness of the biological networks and also identifies the key players in biological processes. A correlation between a node's structural importance in the network and its functional importance commonly referred as centrality-lethality rule is well understood using centrality concepts which were extensively studied. In biological networks, the high centralityproteins are likely to be coded by the essential genes. Centrality measures such as degree, closeness and betweenness measures aids in the identification of the nodes that correlates with gene essentiality [19]. The centrality measures taken individually capture different aspects of gene essentiality, but the combination of them yielded more accurate predictions than using only one of the measures. In our work we make use of method developed by Manimaran *et al* to identify the conditional essentiality of genes in human based on subnetworks of only those genes that are expressed under defined conditions [13]. In our earlier work, we have utilized the binary classification of Support vector machine (SVM) to identify the essential genes in *E coli* where both the positive and negative data sets were available. With the lack of availability of negative data sets in our study we make use of one class SVM for predicting the essential genes in human. The predicted essential genes further aids in identification of potential drug targets.

2. Materials and methods

2.1 Centrality measures

The Human protein–protein interaction data was collected from the Human Protein Reference Database,HPRD which is a resource for experimentally derived manually curated scientific information about the human proteome including protein–protein interactions, post-translational modifications (PTMs) and tissue expression [20].The Human Protein-protein interaction (HPPI) network consists of 9,617 proteins with 39,240 edges. The interactions of HPPI were analyzed using the following three principalcentrality measures namely degree centrality, closeness centrality and betweenness centrality.The mathematical form of the centrality measures are asfollows (for N nodes)

The degree centralityis based on the idea that important nodes are those with the largest number of edges to other nodes in the graph. The degree centrality of a node *i* is defined as

$$C_i^D = \frac{k_i}{N-1} = \frac{\sum_{j \in G} a_{ij}}{N-1}$$

Where $k_i$is the degree of node *i*.

The closeness centralityof a node *i* is based on the concept of minimum distance or geodesic d_{ij} , i.e. the minimum number of edges traversed to get from *i* to *j* and is defined as

$$C_i^c = (L_i)^{-1} = \frac{N-1}{\sum_{j \in G} d_{ij}}$$

Where L_i is the average distance from *i* to all the other nodes.

The betweenness centrality, in its basic version, is defined by assuming that the communication travels just along the geodesic. If n_{jk} is the number of geodesics linking two nodes jand *k*, and $n_{jk}(i)$ is the number of geodesics linking the two nodes *j* and *k* that containnode *i*, the betweenness centrality of node *i* can be defined as

$$C_i^B = \frac{1}{(N-1)(N-2)} \sum_{j \epsilon G, j \neq i} \sum_{k \epsilon G, k \neq i. k \neq j} n_{jk}(i) \Big/ n_{jk}$$

2.2 Machine learning approach

Machine learning can be defined as a as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The SVM algorithm generally is a binary-class algorithm and requires both positive and negative datasets. But if there exits only one dataset i.e positive data set in such scenarios One-Class SVM should be considered. One-Class SVM algorithm maps the data into a feature space H using an appropriate kernel function, and then trying to separate the mapped vectors from the origin with maximum margin. In our work we have used one class SVM using the LIBSVM package [21-23]. The scores of centrality measures were considered as input features for training and these input vectors were trained Radial Basis Function kernel. The optimal kernel parameters, cost C (10) and gamma G (0.0001), were obtained through grid search and the dataset trained with five-fold cross validation.

The function to predict output using one class SVM is discussed as follows

.Let x_1, x_2, $x_{3\ldots}$.$x_l$be training samples belonging toone known class X, where X is a compact subset of R^N . Let Ø: X→H be a kernel map which transforms thetraining samples to another space. Then, to separate the dataset from the origin, one needs to solve the following quadraticprogramming problem:

$$\min \frac{1}{2}||w||^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \rho$$

S.t. w Ø (x_i) ≥ ρ - ξ_i , i=1, 2, 3…..l,$\xi_i \geq 0$

Nonzero slack variables ξ_i are penalized in the objective function. The decision function corresponding to w and ρ is

$$f(x) = \text{W Ø (x)} - \rho$$

The above equation will be positive for most samples x_i contained in the training set. V∈ (0, 1) is a parameter which controlsthe number of samples contained in the hyper sphere.

3. Results and discussion

We have constructed the Human protein –protein interaction network (HPPI) from the data collected from the HPRD database. After curation of the data sets, removal of self and palindromic interactions and removal of CGP islands the core Human PPI network is comprises of 9204 nodes with 36726 edges. The topological properties were also being calculated for the constructed HIPPI network the scaling exponent was observed to be 1.89 which indicates that the constructed network is scale free network and he same is depicted in **Figure 1**. Also the maximum degree and average degree were observed to be 296 and 14 respectively.

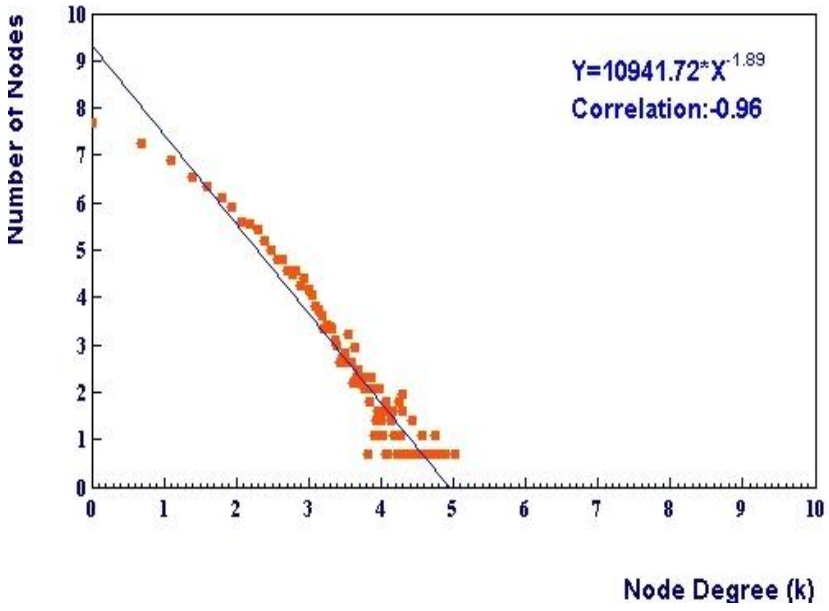


Figure 1: The degree distribution of the HPPI network was observed to follow powerlaw with an exponent of $\gamma = 1.89$, exhibiting the scale-free nature.

Among all the properties graph centrality measures aids in identification nodes that are functionally crucial/central in the network by ranking elements of a network. Identification of central nodes in biological networks paves way for delivering new hypotheses which in turn lead to invention of more rational approaches in experimental design. Different centrality measures scores and ranks the nodes based on different concepts. In case of protein-protein interaction network the centrality of a protein correlates with its essentiality. In our work we have used three

different centrality measures degree, closeness and betweenness for predicting the essential genes in humans. Degree centrality emphasizes that a central node is involved in a large number of interactions. Closeness Centrality indicates important nodes that can communicate quickly withother nodes of the network. Betweenness Centrality shows that nodes which are intermediate between neighbors rank higher. Without these nodes, there would be no way for two neighbors to communicatewith each other. Thus, betweenness centrality shows important nodes that lie on a high proportion of paths between other nodes in the network [24]. The three centrality scores for each of the node were calculated for 9204 proteins of core network. The proteins were then ranked based on their centrality scores and compared the lists with 1950 known human essentialgenes. Interestingly, the majority of the essentialgenes were among those in top of any of the threecentrality lists, whereas only a few essential genes were found in the bottom. In order to nullify the deprivation of any essential gene that occurred in the bottom in terms of the rankingwe attempted to combine the threefeatures using one class SVM algorithm to predict gene essentiality based on combinations of the threecentrality measures.Thefeatures selected were the three network centrality measures,namely, degree centrality (DC), closeness centrality (CC) andbetweenness centrality (BC). The training data consisted of only a positive data set comprising of centrality measures for the 1950 known essential genes. Using this model, when predictions were made 3786 genes were predicted to be essential which implies that about 41% of the genes in HPPI network were found to be essential. A fivefold cross validation with accuracy 79.23% was obtained and 854 genes were found to be in common with 1950 known essential genes that served as positive training dataset. Also we have tried to corroborate the gene essentiality of the predicted genes through a literature survey. The annotation of some of the top ranking genes among the predicted essential genes for the vindication of their essentiality as indirectly provided by NCBI gene resource is as follows

C3 known as Complement component C3 plays a central role in the activation of complement system. Its activation is required for both classical and alternative complement activation pathwaysproving its essentiality. Kininogen 1(***KNG1***) uses alternative splicing to generate two different proteins- high molecular weight kininogen (HMWK) and low molecular weight kininogen (LMWK) of which HMWK is essential for blood coagulation and assembly of the kallikrein-kinin system.The protein encoded plasminogen (***PLG***) is a secreted blood zymogen that is activated by proteolysis and converted to plasmin and angiostatin. Plasmin dissolves fibrin

in blood clots and is an important protease in many other cellular processes while angiostatin inhibits angiogenesis and thus proves the essentiality of PLG. ***SKP1***, S-phase kinase-associated protein 1 encodes a component of SCF complexes, which are involved in the regulated ubiquitination of specific protein substrates, which targets them for degradation by the proteasome. ***TYROBP*** known as TYRO protein tyrosine kinase binding protein encodes a transmembrane signaling polypeptide which may associate with the killer-cell inhibitory receptor (KIR) family and may act as an activating signal transduction element. The Golgi membrane protein***GOLM***1 encodes type II Golgi transmembrane protein which processes proteins synthesized in the rough endoplasmic reticulum and assists in the transport of protein cargo through the Golgi apparatus.

Similarly the ***MMP2*** matrix metallopeptidase 2 is involved in the breakdown of extracellular matrix in normal physiological processes, such as embryonic development, reproduction, and tissue remodeling, as well as in disease processes, such as arthritis and metastasis. The enzyme plays a role in endometrial menstrual breakdown, regulation of vascularization and the inflammatory response. The peroxisomal biogenesis factor 19***PEX19*** is necessary for early peroxisomal biogenesis. It acts both as a cytosolic chaperone and as an import receptor for peroxisomal membrane proteins (PMPs) which are essential for the assembly of functional peroxisomes. The gene ***APOA1*** encodes Apo lipoprotein A-I, which is the major protein component of high density lipoprotein (HDL) in plasma. The protein promotes cholesterol efflux from tissues to the liver for excretion, and it is a cofactor for lecithin cholesterolacyltransferase (LCAT) which is responsible for the formation of most plasma cholesteryl ester.***KCNB1*** , potassium voltage-gated channel, Shab-related subfamily, member 1diverse functions include regulating neurotransmitter release, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction, and cell volume and hence proved its essentiality [25] . Thus our approach utilizing the integrative graph theory and machine learning algorithms has efficiently prioritized the essential genes in human.

4. Conclusion

The huge availability of biological data makes the analysis of biological systems more complex.Understanding these complex systems requires a bottom up approach i.e. breaking the complex system into small elementary constituents and mapping out the interactions between

these components, can be characterized as network.The mathematical discipline which underpins the study of these complex biological networks is graph theory. Our work prioritizes the importance of applications graph theoretical approaches on biological networks by identification of essential genes in human protein- protein interaction network through network analysis.

5. Acknowledgement

The authors TK, SBR and PM would like to thank Department of Science and Technology, Government of India, for their financial support (DST-CMS GoI Project No. SR/S4/MS: 516/07 Dated 21.04.2008).

6. References

1. Erdos, P., and Renyi A., on random graphs,*IPubl. Math. Debrecen***6** (1959), 290-297.
2. Watts, Duncan J., and Steven H. Strogatz. Collective dynamics of small-world networks, *Nature* **393** (1998), 440-442.
3. Albert, R., and Barabási, A. L., Statistical mechanics of complex networks. *Reviews of modern physics*, **74** (2002), 47.
4. Wasserman, Stanley, *Social network analysis: Methods and applications*, Cambridge university press, **8** (1994).
5. Scott, John, *Social network analysis: A handbook,* London: Sage. (2000).
6. Strogatz, Steven H, Exploring complex networks, *Nature*, **410** (2001), 268-276.
7. Dorogovtesev, Sergei N., and José FF Mendes, *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, (2013).
8. Barabási, Albert-László, and Réka Albert., Emergence of scaling in random networks, *science* **286** (1999), 509-512.
9. Holme, Petter, Mikael Huss, and HawoongJeong, Subnetwork hierarchies of biochemical pathways, *Bioinformatics*, **19** (2003), 532-538.
10. Wuchty, Stefan, and Peter F. Stadler, Centers of complex networks, *J. Theor Biol*, **223** (2003), 45-53.
11. Newman, M. E, Modularity and community structure in networks, *Proc. Natl. Acad. Sci. U. S. A.*, **103** (2006), 8577-8582.

12. Hegde, S, Manimaran, P andMande, S. C., Dynamic changes in protein functional linkage networks revealed by integration with gene expression data, *PLoS Comput. Biol*., **4** (2008), e1000237.
13. Manimaran, P, Hegde, S, and Mande, S. C., Prediction of conditional gene essentiality through graph theoretical analysis of genome-wide functional linkages, *Mol Biosyst*., **5** (2009), 1936-1942.
14. Koschützki, Dirk, and Falk Schreiber, *Comparison of Centralities for Biological Networks*, Proc. German Conf Bioinformatics, (2004).
15. Bavelas, Alex,A mathematical model for group structures, *Human organization,***7** (1948), 16-30.
16. Freeman, Linton C., Centrality in social networks conceptual clarification, *Social networks*, **1** (1979), 215-239.
17. Latora, Vito, and Massimo Marchiori, A measure of centrality based on network efficiency,*New J. Phys.*, **9** (2007), 188.
18. Latora, Vito, and Massimo Marchiori, How the science of complex networks can help developing strategies against terrorism,*Chaos, solitons & fractals* **20** (2004), 69-75.
19. Jeong, Hawoong, et al., Lethality and centrality in protein networks, Nature, **411** (2001), 41-42.
20. Prasad, TS Keshava, et al., Human protein reference database—2009 update, *Nucleic Acids Res*, **37** (2009), D767-D772.
21. Schölkopf, Bernhard, et al., Estimating the support of a high-dimensional distribution, *Neural Comput*, **13**, (2001), 1443-1471.
22. Fu, Keren, et al., *One-class SVM assisted accurate tracking., Distributed Smart Cameras (ICDSC),* 2012 Sixth International Conference on. IEEE, (2012).
23. Chang, Chih-Chung, and Chih-Jen Lin, LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST),* **2** (2011), 27.
24. Pavlopoulos, Georgios A., et al., Using graph theory to analyze biological networks, *BioDatamining*, **4** (2011), 10.
25. www.ncbi.nlm.nih.gov/gene.