Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS
ISSN 2454-6348

Original Research Article

DOI: 10.26479/2018.0406.05

INFERRING PROTEIN INTERACTION NETWORK OF *MYCOBACTERIUM TUBERCULOSIS H37RV* USING SEQUENCE INFORMATION

Dhammapal Bharne, Damuka Naresh, Vaibhav Vindal[*]

Department of Biotechnology and Bioinformatics, School of Life Sciences,
University of Hyderabad, Hyderabad, India.

ABSTRACT: Protein interaction network helps to understand the general mechanism behind the complex biological systems which in turn provide insights into disease mechanism and pathways. It can be modulated towards drug discovery for the infectious organism. With vast sequence data availability, computational methods play a major role to infer protein-protein interaction. The present study explores sequence-based information to infer protein interaction network in *Mycobacterium tuberculosis*, a causative agent of one of the leading infectious diseases Tuberculosis. A model was built using support vector machine (SVM) based classification through amino acid conjoint triad from interacting protein pairs from the DIP database. 162,528 protein interaction pairs of *M. tuberculosis H37Rv* were identified through the model. It was observed that 53,602 protein pairs were already known to be interacting or co-expressing in multiple microarray experiments. These protein pairs were considered as significant interaction pairs. A protein interaction network built for core interactions had 53,424 edges connected across 1,368 nodes. The highest degree was observed for pknB, a serine/threonine-protein kinase which may serve as a potential drug target for tuberculosis. Further, two conserved hypothetical proteins, Rv3879c and Rv3909, were found to be hub proteins. Exploration of such hubs will assist in understanding the regulation of genes and disease processes which may lead to develop better intervention strategies for the disease.

KEYWORDS: Support vector machine, conjoint triad, protein interaction network, degree, hub.

Corresponding Author: Dr. Vaibhav Vindal* Ph.D.

Department of Biotechnology and Bioinformatics, School of Life Sciences,
University of Hyderabad, India. Email Address: vaibhav@uohyd.ac.in

Peer review under responsibility of Life Science Informatics Publications

1. INTRODUCTION

Now is the time for modern biology to switch from the study of single molecules to network-based systems. The functions of various cellular processes are determined by protein interactions rather than an individual protein. Though, a large number of protein-protein interactions (PPIs) are predicted with experimental methods, it covers only a fraction of total protein-protein interactions [1, 2]. In this regard, computational methods play a vital role in achieving the complete protein interaction network [3, 4, 5]. Various in silico methods are proposed to identify PPIs such as phylogenetic profile, gene operon and domain-based methods [6, 7, 8]. Recently, it is reported that the identification of PPIs based on only sequence information is more universal [9, 10]. Amino acid conjoint triad method identifies PPIs using only protein sequence information [10]. The sequence based method is powerful for identifying protein-protein interactions and in exploring the networks for newly discovered proteins with unknown biological functions. In spite of large research efforts, tuberculosis is still one of the leading infectious diseases in the world [11, 12]. Multidrug resistance and HIV co-infections provoke to investigate a novel system to combat the disease [13]. A protein interaction network of *M. tuberculosis H37Rv* helps in understanding cellular physiology and also identifying suitable drug targets [14]. The present study employs amino acid conjoint triad method to identify interacting protein pairs in *M. tuberculosis H37Rv*. Since the interaction network is a large scale real world graphical data, it is an ideal challenge for bioinformatics research.

2. MATERIALS AND METHODS

1. Dataset Preparation

Protein interaction data for *M. tuberculosis H37Rv* was downloaded from DIP databases [15]. These interacting pairs were used to prepare positive and negative training examples for conjoint triad method [10]. During this process, 343 triads of amino acids with similar physiochemical roles were generated. The frequency of each triad was calculated and then normalized to represent a protein sequence. An interaction pair was obtained by concatenating two proteins with normalized triad frequencies. Therefore, each interaction pair had 686 conjoint triads of amino acids. Since the protein interactions are symmetrical, a reverse directional calculation was also employed. In order to prepare negative training examples, Euclidean distances among the proteins from the interaction pairs were calculated. Protein pairs with Euclidean distance value above the average of smallest and largest Euclidean distance values were used as negative training examples.

2. Generation of SVM Model

In order to generate a SVM model, radial basis function (RBF) from LIBSVM software [16] was employed. Positive and negative training examples were scaled together with default parameters. The scaled data was then randomized 1000 times and grid search was performed to identify the best C and γ parameter values for the RBF kernel. The scaled data was trained using the best C and γ values to generate a SVM model.

3. Prediction of PPIs

Complete proteome set of *M. tuberculosis H37Rv* was downloaded from NCBI genome [https://www.ncbi.nlm.nih.gov/genome/]. Conjoint triad frequencies were calculated and normalized for each of the protein sequences. Protein pairs were generated by concatenating every protein sequence with every other protein sequence. If a protein pair was predicted positive by the built SVM model, it was considered as interacting pairs.

4. Inferring Protein interaction map

Predicted protein interaction pairs were compared with interaction data from the STRING database [17, 18] and the MPIDB database [19]. They were also compared with already known interactions from Wang *et al*., [20] and Liu *et al*., [21] data. Further, predicted interaction pairs were validated using co-expression analysis [22]. During this process, microarray experiments from NCBI GEO [https://www.ncbi.nlm.nih.gov/geo/] were filtered with "*Mycobacterium tuberculosis*" and a sample size of at least 10. Further, only experiments with ORF names were considered. The experimental data was log base 2 transformed and imputed with k-nearest neighbor method. It was quantile normalized to remove sources of variations. Expression profiles of the proteins in the predicted protein interactions were extracted from the normalized data and Pearson correlation coefficients were calculated. The pairs of proteins with correlation value of at least 0.8 in more than one microarray experiment were considered as co-expressing proteins pairs [22]. Predicted protein pairs supported by interaction data from the STRING database, Wang *et al.,* data, Liu *et al.,* data or co-expressing in multiple microarray experiments were considered as significant protein interaction pairs. An interaction map was generated for these significant pairs using VisANT software [23, 24]. High degree nodes and clustering coefficients were derived from the interaction network.

5. Functional Enrichment

Top 5 percent high degree nodes of the protein interaction network were considered as hubs. These hubs were used to find their functional enrichment in ontologies such as biological process, cellular component and molecular function through PANTHER over-representation test [25]. P-value cut off is set to 0.05 in order to obtained significant ontological terms. Bonferroni correction for multiple testing was also considered during the analysis.

3. RESULTS AND DISCUSSION

1. Protein-protein interactions

It was observed that the training data set had 38 positive examples and 444 negative examples. The grid search of the scaled data generated the best C value of 8.0 and γ value of 0.0078125. A SVM model generated using these values through RBF kernel was used to predict protein interactions in *M. tuberculosis H37Rv*. In the present study, a total of 162,528 PPIs were predicted using the built SVM model. It was observed that the predicted interactions were scattered among 1766 proteins. However, it leads to the coverage of 45.12 percent of total proteins of *M. tuberculosis H37Rv*.

2. Known Interacting Pairs

It was observed that some of the predicted PPIs were already reported in the existing literature such as Wang et al. and Liu et al. data. Some of them were also available in the databases such as STRING database and MPIDB database. Further, 75 microarray experiments were found to have ORF names in expression data. Co-expression analysis of microarray experiments for the proteins in the predicted PPIs was also performed. The Table 1 represents the number of predicted PPIs that overlap with existing literature and databases as well as co-expresses in multiple microarray experiments. It is clear from the table that 59,019 PPIs overlaps with existing knowledge. It was found to have 53,602 unique PPIs. Therefore, more than 32.98 percent of the predicted PPIs were supported from the know data. These predicted PPIs were considered as significant proteins interaction pairs in the present study.

Table 1: Overlapping PPIs

Sr. No.	Source	Dataset Size	Overlapped PPIs
1	STRING database	796,610	15,433
2	MPIDB database	19	5
3	Liu et al.	43,136	1,304
4	Wang et al.	8,242	60
5	Co-expression analysis	1,705,422	42,217
		Total PPIs	**59,019**

3. Protein Interaction Network

Significant protein interaction pairs were used to construct a protein interaction map of *M. tuberculosis H37Rv*. It was observed that there were 183 isolated nodes with a degree of 1. These nodes were removed to get a core interaction network. The Figure 1 represents an interaction map generated from the core interaction network using the VisANT software. The network has 53,424 black colored lines called edges representing core protein interactions and 1,368 green colored boxes called nodes representing the interacting proteins. Visualization of the interaction network indicates that it is a simple network. Degree distribution follows the power law; hence the interaction network is a scale-free network [26]. The highest degree was observed for Rv0014 (pknB) with the value of 871 followed by Rv2524c (fas) with the value of 663 and Rv2379c (mbtF) with the value of 648. When the degree threshold was set at 50 and above, 608 hub nodes were identified. Correlation of node degree with clustering coefficient was found to be 0.44 indicating that it is a well clustered network. The average of clustering coefficients of the nodes was found to

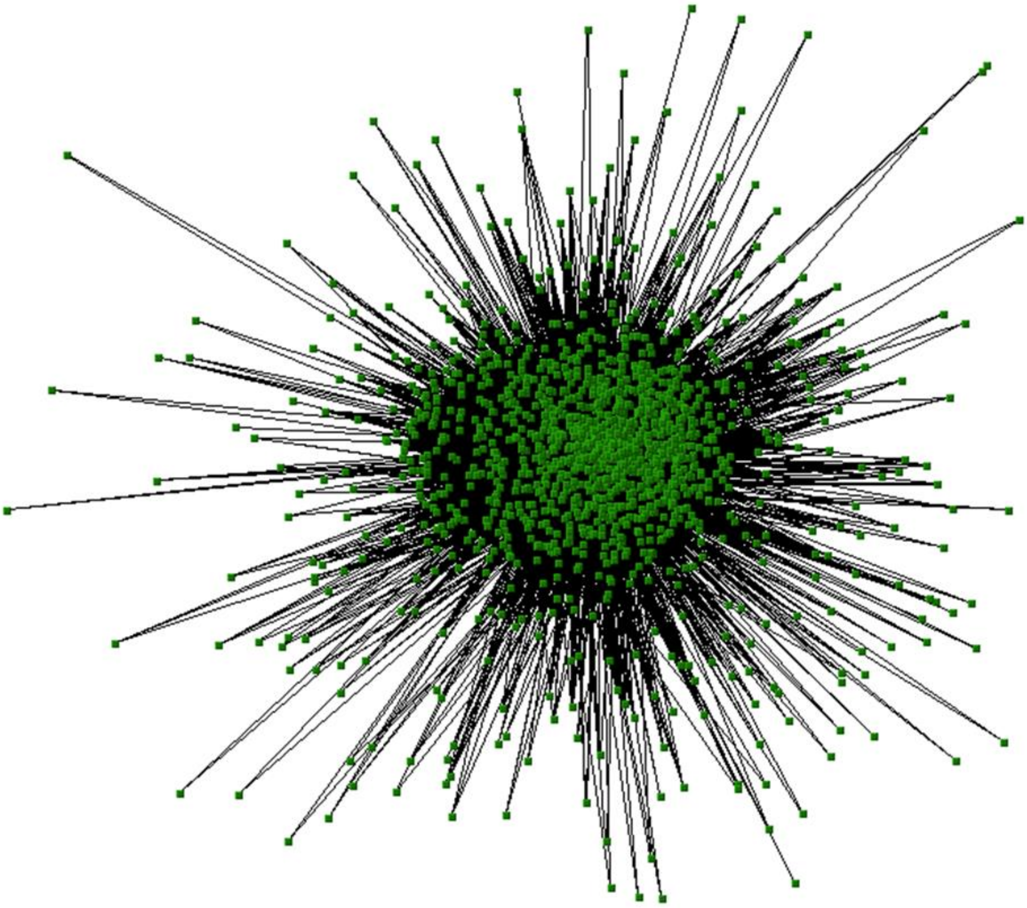

Figure 1: Protein Interaction network of *M. tuberculosis H37Rv*

be 0.459 suggesting that there are many nodes which were well clustered. There were 30 fully connected subgraphs. These subgraphs represent related functions of the protein and strongly suggest that they could form a complex [27]. As the network is a scale-free network, deletion of high degree nodes will destroy the interaction network significantly. Therefore, they can be exploited as potential targets for effective control and cure of the disease.

4. Functional annotation

Top 5 percent of hubs were employed for functional annotation [28, 29]. It was observed the cutoff degree to be 378 which was possessed by three nodes, viz., Rv2447, Rv3879c and Rv2931. Thus value 5 percent of hubs constituted 33 nodes. Functional annotation of these nodes was performed using PANTHER [16] over-representation test. It indicated that most of the hub proteins were involved in fatty acids, amino acids and lipid metabolic processes. Further, they were mostly enriched in transferase, hydrolase and ligase activities.

5. Functional categories

Top 5 percent hub proteins were categories based on Tuberculist [30] data. The Figure 2 indicates the proportion of hubs under different categories. It is clear from the figure that most of the hubs are involved in lipid metabolism, cell wall and cell processes. Further, it was observed that Rv3879c and Rv3909 the two conserved hypothetical proteins which are also hubs. Further, it was observed that 16 hubs are essential genes as observed from the Database of Essential Genes [31, 32].

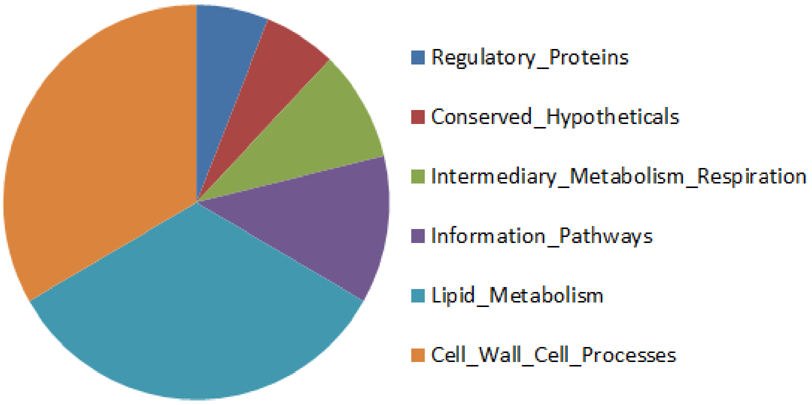

Figure 2: Tuberculist categories of Hubs

4. CONCLUSION

In the present study, support vector machine based conjoint triad was efficiently employed to predict PPIs of *M. tuberculosis H37Rv*. Using this approach, 162,528 interacting pairs were identified. It was observed that more than 32.98 percent of predicted PPIs overlap with already known PPIs. Therefore, this approach is effective to identify PPIs using only sequence information. The proteome coverage of predicted PPIs was 45.12. This suggests that different approaches can be employed to predict number of interacting pairs in order to achieve a complete protein network of *M. tuberculosis H37Rv*. Hubs were identified which may serve as potential drug targets for tuberculosis. Further, conserved hub hypothetical proteins were also found which can be explored further to understand their potential role in cellular physiology and disease mechanism of the organism.

ACKNOWLEDGEMENT

The Departmental facilities were supported by grants from UPE-II and DST-PURSE. DB is supported by UGC fellowship. Authors acknowledge computational support from Bioinformatics Infrastructure Facility, University of Hyderabad, India.

CONFLICT OF INTEREST

None.

REFERENCES

1. Zhou H, Wong L. Comparative analysis and assessment of *M. tuberculosis H37Rv* protein-protein interaction datasets. BMC Genomics. 2011; 12(3):S20.
2. Han J-DJ, Dupuy D, Bertin N, Cusick ME, Vidal M. Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol. 2005; 23(7):839-844.
3. Wodak SJ, Méndez R. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol. 2004; 14(2):242-249.
4. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol. 2002; 12(3):368-373.
5. Raman K. Construction and analysis of protein–protein interaction networks. Autom Exp. 2010;

2(1):2.

6. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999; 96(8):4285-4288.
7. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature. 1999; 402(6757):86-90.
8. Jothi R, Cherukuri PF, Tasneem A, Przytycka TM. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. J Mol Biol. 2006; 362(4):861-875.
9. Skrabanek L, Saini HK, Bader GD, Enright AJ. Computational prediction of protein–protein interactions. Mol Biotechnol. 2008; 38(1):1-17.
10. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci U S A. 2007; 104(11):4337-4341.
11. Hingley-Wilson SM, Sambandamurthy VK, Jacobs Jr WR. Survival perspectives from the world's most successful pathogen, *Mycobacterium tuberculosis*. Nat Immunol. 2003; 4(10):949-955.
12. Raviglione M, Sulis G. Tuberculosis 2015: burden, challenges and strategy for control and elimination. Infect Dis Rep. 2016; 8(2):6570.
13. Leibert E, Danckers M, Rom WN. New drugs to treat multidrug-resistant tuberculosis: the case for bedaquiline. Ther Clin Risk Manag. 2014; 10:597-602.
14. Raman K, Yeturu K, Chandra N. targetTB: a target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. BMC Syst Biol. 2008; 2:109.
15. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S-M, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002; 30(1):303-305.
16. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2011; 2(3):27.
17. Mering Cv, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 2003; 31(1):258-261.
18. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005; 33:D433-D437.
19. Goll J, Rajagopala SV, Shiau SC, Wu H, Lamb BT, Uetz P. MPIDB: the microbial protein interaction database. Bioinformatics. 2008; 24(15):1743-1744.
20. Wang Y, Cui T, Zhang C, Yang M, Huang Y, Li W, et al. Global protein- protein interaction

network in the human pathogen *Mycobacterium tuberculosis H37Rv*. J Proteome Res. 2010; 9(12):6665-6677.

21. Liu ZP, Wang J, Qiu YQ, Leung RK, Zhang XS, Tsui SK, et al. Inferring a protein interaction map of *Mycobacterium tuberculosis* based on sequences and interologs. BMC Bioinformatics; 2012; 13:S6.
22. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res. 2004; 14(6):1085-1094.
23. Hu Z, Mellor J, Wu J, DeLisi C. VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinformatics. 2004; 5(1):17.
24. Granger BR, Chang YC, Wang Y, DeLisi C, Segre D, Hu Z. Visualization of metabolic interaction networks in microbial communities using VisANT 5.0. PLoS Comput Biol. 2016; 12(4):e1004875.
25. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003; 13(9):2129-2141.
26. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999; 286(5439):509-512.
27. Langfelder P, Horvath S. Fast R functions for robust correlations and hierarchical clustering. J Stat Softw. 2012; 46(11):i11.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genetics. 2000; 25(1):25-29.
29. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic Acids Res. 2015; 44(D1):D336-D342.
30. Lew JM, Mao C, Shukla M, Warren A, Will R, Kuznetsov D, et al. Database resources for the tuberculosis community. Tuberculosis (Edinb). 2013; 93(1):12-17.
31. Zhang R, Ou HY, Zhang CT. DEG: a database of essential genes. Nucleic Acids Res. 2004; 32:D271-D272.
32. Luo H, Lin Y, Gao F, Zhang CT, Zhang R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res. 2013; 42(D1):D574-D580.