OXFORD UNIVERSITY PRESS | Briefings in Bioinformatics

Computational analysis of protein interaction networks for infectious diseases

Journal:	*Briefings in Bioinformatics*
Manuscript ID:	BIB-15-0058.R2
Manuscript Type:	Paper
Date Submitted by the Author:	08-Jul-2015
Complete List of Authors:	Pan, Archana; Pondicherry University, Centre for Bioinformatics Lahiri, Chandrajit; Sunway University, Department of Biological Sciences Rajendiran, Anjana; Pondicherry University, Centre for Bioinformatics Shanmugham, Buvaneswari; Pondicherry University, Centre for Bioinformatics
Keywords:	Protein interaction network, Infectious disease, Pathogen, Computational analyses, Centrality, Modularity

Computational analysis of protein interaction networks for infectious diseases

Authors: Archana Pan[1#*], Chandrajit Lahiri[2,#], Anjana Rajendiran[1] and Buvaneswari Shanmugham[1]

[1]Center for *Bioinformatics*, School of Life Sciences, Pondicherry University, Puducherry – 605014, India

[2]Department of Biological Sciences, Sunway University, 47500 Bandar Sunway, Selangor, Malaysia

#These authors contributed equally to this work

***Corresponding Author:** Archana Pan, **E-mail :** archana@bicpu.edu.in

Phone: +91 413 2654584

Fax: +91 413 2655211

Authors: Archana Pan, **E-mail :** archana@bicpu.edu.in

Chandrajit Lahiri, E-mail **:** chandrajithlahiri@gmail.com

Anjana Rajendiran, Email: anjana@mails.bicpu.edu.in

Buvaneswari Shanmugham, Email: buvanisuriya@bicpu.edu.in

Abstract

Infectious diseases caused by pathogens, including viruses, bacteria and parasites, pose a serious threat to human health worldwide. Frequent changes in the pattern of infection mechanisms and the emergence of multidrug resistant strains among pathogens have weakened the current treatment regimen. This necessitates the development of new therapeutic interventions to prevent and control such diseases. To cater to the need, analysis of protein interaction networks (PINs) has gained importance as one of the promising strategies. The present review aims to discuss various computational approaches to analyse the PINs in context to infectious diseases. Topology and modularity analysis of the network with their biological relevance, and the scenario till date about host-pathogen and intra-pathogenic protein interaction studies were delineated. This would provide useful insights to the research community thereby enabling them to design novel biomedicine against such infectious diseases.

Keywords: Protein interaction network, Infectious disease, Pathogen, Computational analyses, Centrality, Modularity.

Author Biography

Archana Pan, PhD in Bioinformatics, currently serves as an Assistant Professor at Centre for Bioinformatics, Pondicherry University, Pondicherry, India. Her research interests include comparative genomics, molecular evolution and drug design.

Chandrajit Lahiri, PhD in Molecular Microbiology, is currently a Senior Lecturer at Department of Biological Sciences, Sunway University, Selangor, Malaysia. His research interests encompass Systems Biology and Evolutionary Bioinformatics.

Anjana Rajendiran is registered for PhD at Centre for Bioinformatics, Pondicherry University, Pondicherry, India. Her research interests include comparative genomics, machine learning, and microRNA prediction.

Buvaneswari Shanmugham is registered for PhD at Centre for Bioinformatics, Pondicherry University, Pondicherry, India. Her research interests lie in comparative genomics, structural bioinformatics and genetic algorithms.

Key Points

1. Infectious diseases have posed serious health concerns worldwide
2. Conventional approaches have become almost ineffective in dealing with the issue
3. Non-conventional computational approach entailing protein interaction network analysis has gained importance to give meaningful directions
4. Topological and Modularity analyses of PINs can be employed by researchers to obtain essential proteins as key therapeutic targets
5. Analyses involving these would pave the way to succeed in generating novel biomedicines

Introduction

Infectious diseases have been threatening human population since time immemorial. These have become ever-increasing worldwide public health concern, with parasitic, bacterial and viral diseases, representing more than half of the leading causes of morbidity and mortality. While viral influenza vaccines are to be reformulated annually, several other viral infectious diseases, such as those from hepatitis C and HIV-1 are a cause of panic since decades. Parasitic diseases like malaria and multidrug resistant bacterial strains of *Mycobacterium* and *Salmonella* are on-going pandemics. Different emerging infectious diseases *viz.*, nosocomial infections caused by *Acinetobacter*, swine H1N1 influenza, avian H5N1 influenza, severe acute respiratory syndrome (SARS), and dengue fever have posed themselves to be new constant threats [1].

To deal with these severe pathogenic threats, several health intervention strategies have been undertaken. However, the prospects for finding new vaccines or antibiotics against such pathogens are especially poor. This is due to the ever-changing mechanism of infecting the host as in the cases of viruses [1]. It might also be due to the blockades provided by the outer membrane to the entry of some existing antibiotics in case of gram negative bacteria [2]. Thus, it is quite evident that the conventional strategies for dealing with such deadly pathogens would be less effective or ineffective completely, to emerge victorious against their strategies to evade therapeutic interventions. In such cases, the complexities posed can be solved by adopting some non-conventional computational approaches.

Over the last few decades, biologists understood gradually that a set of complex interactions between the numerous constituents of a cell, gives rise to different biological phenotypes. Amongst these, proteins, being the functional unit of the cell of any living organism, always act

in unison with others to achieve specific functional goals *viz*., transcriptional activation/repression; immune, endocrine, and pharmacological signaling; cell-to-cell interactions; and metabolic and developmental control [3]. These protein-protein interactions (PPI) lead to a mosaic mesh or network of interactions, commonly known as protein interaction networks (PINs). Analyses of such PINs are increasingly serving as the non-conventional approach to understand the complexity of infectious diseases. However, the augmentation of the PINs, created from high-throughput experimental and/or computational data, has necessitated effective analytical techniques for those networks, to be used to unravel the molecular basis of the aforementioned infectious diseases. The current review entails different computational approaches for analysing protein interaction networks expected to be involved in the interaction mechanism of infection. These might lead to find avenues for the identification of novel targets and render them as systems biomedicines.

The necessity of the generated PIN

With the advances of the post genomic era, there has been an enormous increase in the investigations upon the structure, function and control of the participating proteins as key regulators in diseases. This is because, the identification of a handful of proteins to be targeted is considered as the objective of the whole intervention process. The numbers of proteins, as targets, should always be limited, to improve the efficacy and specificity of a well-defined drug. However, ensuring a limited number of proteins from an array becomes an ever challenging task to the conventional experimentalists. Thus, new approaches, for generating viable candidates as interventional targets for infectious diseases, are need of the hour.

The complexities of any infectious diseases are mainly due to the intricate interactions between sets of proteins involved in the process. Interactions between proteins are visualized by networks created by mapping those complex interactions. These protein interaction networks (PINs) have gradually gained importance in an attempt to address the complexities of the diseases. Such mapping can be done based on a number of experimental data sources including, but not limited to, two-hybrid systems [4], mass spectrometry [5], protein chip technologies [6]. They can also be generated through various computational approaches encompassing genome-based [7, 8], sequence-based [9, 10], structure-based [11, 12] and machine-learning-based techniques [13, 14]. However, analysing these networks, to achieve the ultimate goal of limiting target sets for health intervention, now becomes the most challenging task.

The resources of PPIs

While the high throughput techniques generated interaction data for proteins, initiatives were taken to integrate them and prepare comprehensive open databases for further analyses. There are a number of standardized open sources each having a different style of representing the protein interaction datasets. They are mostly based on the organisms worked upon in detail and of basic interest amongst researchers. Of these, Human Protein Reference Database (HPRD) stores information on human protein interactions, along with protein functions, post-translational modifications (PTMs), enzyme-substrate relationships, and subcellular localization [15]. Sub-categorised HomoMINT [16] arises from the Molecular Interaction Database (MINT) [17] which comprises interactions, inferred from orthologs in model organisms. For the yeast PPI data, special importance has been given in the Biological General Repository for Interaction Database (BioGRID) [18].The current BioGRID release [May 2015, version 3.3.124] lists 287,619 non-

redundant yeast protein interactions thereby making it the largest database for this organism besides more than 30 others, having a total of 574,378 non-redundant interactions. Another such database focusing on yeast data is the Database of Interacting Proteins (DIP) integrating data from the correlation of protein sequence and RNA expression profiles through a carefully curated computational process [19].

Besides the above mentioned focused databases, there are others, having listed the protein interactions from a set of organisms. These are the Munich Information Center for Protein Sequences (MIPS) [20], the Biomolecular Interaction Network Database (BIND) [21], a component of the Biomolecular Object Network Databank (BOND), the Search Tool for Recurring Instances of Neighbouring Genes/Proteins (STRING) [22] and IntAct [23], each having its own uniqueness. MIPS lists a description and the binding regions of interacting partners. BIND highlights the interactions between two or more molecules which form functional molecular complex units and pathways arising from those interacting in a sequence. The STRING database entails both physical and functional associations derived from genomic context, high-throughput experiment, coexpression and previous knowledge. Apart from interaction data, IntAct enlists interactions between DNA, RNA, and small-molecules. Furthermore, some databases including STRING [22], GeneMANIA [24], FunCoup [25] and ConsensusPathDB [26] provide a highly comprehensive data by integrating PPIs from other online resources. STRING imports PPI data from different primary databases, including MINT, HPRD, BioGRID, DIP, BIND, IntAct and PDB. GeneMania provides functionally similar genes for the query gene list along with interactive functional association network utilizing information from GEO, BioGRID, Pathway Commons and I2D. FunCoup includes information on functional couplings between genes and gene products based on gold standards (KEGG, Corum, iRefindex

etc.). ConsensusPathDB enlists a seamless interaction network from different public resources, including BIND, BIOGRID, DIP, IntAct, MINT and MIPS-MPPI. There are different resources providing intra-pathogenic and host-pathogen PPI data, such as PATRIC, PRIMOS, HPIDB, PHI, VirusMentha, VirusHostNet [27-33]. The number of intra-pathogenic and host-pathogen PPIs is tabulated in Supplementary Table 1.

The technicalities of PIN analysis

The databases stated above list the interactions of the proteins from existing empirical and theoretical results. As such, an attempt to construct a network or an interactome, by integrating those interactions, might yield one, which can be random like the one proposed by Erdös and Renyi [34] or a small-world type proposed by Watts and Strogatz [35]. Both these types build up a fairly homogeneous network in which, each node has approximately the same number of links. However, only those interactomes, which strictly follow the power law, are free of a characteristic scale. In these cases, the connectivity distribution, P(k), of a node in a network getting connected to k other nodes, decays exponentially for large values of k. These scale-free networks are essentially the real world networks [36] with a heavy tailed degree distribution. Thus, it is imperative to construct biologically viable real networks, comprising the proteins responsible for the infectious diseases. Their subsequent analyses, in essence, would then lead us to our ultimate goal of identifying important targets for health intervention.

The analyses of the interactomes

An overview of various computational approaches for protein interaction network analysis is illustrated in Figure 1.

Topological analyses to identify an important protein

In order to identify the key proteins in a PIN, the importance of the protein is correlated with the number of its interacting protein partners. This gives rise to the concept of such proteins becoming central to a particular network. This is the most basic concept in terms of biological importance and is defined as the degree centrailty (DC) of the protein in a network of interacting proteins. Indeed, high degree proteins (or hubs) are known to correspond to the essential proteins in a network [37]. However, DC is a local and static metric, as it considers only the directly connected neighbours of a protein in a static state. Thus, DC, being the local property of a protein in the network, does not bring out the importance of the protein on a global scale. To indicate such importance based on a protein's global relevance in the network, researchers resort to other centrality measures. These are Closeness centrality (CC), Betweenness centrality (BC) and Eigenvector centrality (EC) [38]. These four important concepts of centrality measures reportedly have been utilized for biological network analyses [39-41].

It is understood that, being the most basic of the centrality measures, DC generally refers to the protein involved in a large number of interactions in a network. However, these interactions might not be in a sequential order so as to carry out particular functions during the primary stages of infection by a pathogen. Conceptually, CC might take care of this fact as it reflects the protein, which is typically "close" to, and can communicate sequentially with the other proteins in the network. Thus, CC is a measure based on the interacting distance of a protein to all other proteins in a network. It is defined as the reciprocal of the total interacting distance from a protein to all other proteins in the network. Again, in a complex phenotype like virulence in infectious diseases, there might be simultaneous interaction of a protein with others to render different functions at the same time. Thus, an important protein should be typically the one,

which lies on a high proportion of interactions between other proteins in the network. An analysis with BC might bring out this fact. Thus, a better measure compared to the DC and CC would be BC, since it would reflect the importance of the protein with respect to its indispensability as it would form the bridge between important hubs of network thereby becoming important. BC of a protein is defined as the number of shortest interacting paths passing through it. However, the ultimate idea of a protein in a network to be important lies in the fact that it should be connected to other important proteins in the network. EC might come into play in such cases [42]. EC brings out the relative importance of the proteins in the network by weighting the connections to other important proteins compared to those of low importance [43].

It has been observed that topological features like DC and BC have gained much importance in serving as attractive drug targets [36, 44, 45]. However, despite their potential to locate such targets, these measures lack in the specificity and/or selectivity along with the high risk of side effects. These, in turn, result in a high likelihood of causing lethality as determined experimentally in the yeast PIN [46, 47]. As lethality is an undesirable attribute in most of the drug discovery applications [3], an alternative measure for betweenness can be thought of. This is known as bridging centrality and proteins with high bridging centrality mainly serve as bottlenecks between two modules. This has been shown to be less lethal, with a value of 34% compared to 42 for BC and 48 for DC in case of yeast PIN [48].

There are other topological properties which have been utilized to measure the compactness and reachability amongst the interacting proteins in the network. One of these is the average path length (APL) which determines the mean of the lengths of the shortest paths between all protein

components of the network [49-51]. The other is network diameter which measures the longest distance between two constituent components [49].

Network decomposition to identify set of important proteins

In general, the PIN for an infectious disease would be on a large scale. Thus, as discussed in the above section, a focus to target just one protein, for therapeutic health intervention, may be of less importance. This might necessitate a decomposition of those large networks to a core of highly interacting proteins through the k-core analysis approach [52].This essentially peels off the proteins connected at the edges, gradually, until the innermost core is reached. After this core, a step further decomposes the network, thereby making this the innermost core with highly connected proteins, interacting with each other. Thus, they can be considered to be the most important ones [42].

Technically speaking, the *k*-core of the graph G is obtained by recursively removing all the vertices of degree less than *k*, until all vertices in the remaining graph have at least degree *k*, by which the complex network can be decomposed [53, 54].

Modularity analyses and functional annotation of clustered proteins

The concept of *k*-core, as discussed above, is one of the metrics to determine the modularity of a network. A modular network groups the components on the basis of their common properties to bring out significant underlying principles. Analyses of these networks become increasingly useful for PINs. This is due to the biological phenomenon of proteins aggregating into complexes, rendering them as functional modules which unify the cohesive components of a molecular function. The identification of such highly correlated functional modules of proteins

can be done by clustering analyses. These protein modules from one species can then be utilised to rationally map and thereby annotate the unannotated proteins in other related genomes.

Besides *k*-core, the clustering techniques can identify cliques. A clique is an induced complete subnetwork where each component vertex is connected to each other. This gives rise to a clustering coefficient of 1 for each of the component vertices. Parametric indices like maximum clique centrality (MCC), maximum neighbourhood centrality (MNC) and density of maximum neighbourhood centrality (DMNC) are offshoots of these concepts and has been utilised lately [42]. Such densely connected subnetworks are expected to form functional units to carry out unique biological processes.

While such density based traditional clustering method is in good practice amongst researchers, new approaches through non-traditional methods have started gaining importance. This is because of their ability to analyse the modularity of the PPI networks with more accuracy. These include the graph-theoretic, topology-based, flow-based, statistical, and domain knowledge-based approaches (data fusion, GO integration) besides the distance-based methods [3]. Of these, the topology- and distance-based modularity analyses focus on the biological distance or similarity between the interacting proteins. Such distance/similarity based matrix can then be utilised to build up the traditional clustering algorithms as in, for instance, Unweighted Pair Group Method with Arithmetic Mean (UPGMA), generally used for calculating evolutionary distance. However, to emerge into more biologically relevant models, instead of only indicating the binary relationships as in the traditional coefficient based ones, sequence similarity, structural similarity and gene expression correlation have started to be used [55-57].

Any attempt to cluster such biologically relevant modular networks would bring out the importance of the interrelationships of the constituent components. To formulate the modularity,

the graph theoretic and the topology-based methods consider the local or global structure of the PPI networks. While the former converts the process of clustering into graph theoretic problems, the latter quantitatively measures the metric features of the networks before formulating the clustering algorithms for modularity analyses. It is to be noted that the graph theoretic features have gained much importance in modularity analyses due to the fact that they can find out the densest subnetworks e.g. Molecular Complex Detection (MCODE), clique percolation. Amongst these, clique percolation method has its advantage of identifying overlapping functional clusters in a typical PPI network. This enables one to detect proteins simultaneously functioning differently in several different modules [3]. The other method of utilising the graph theoretic measures is through partitioning the modular subnetworks, either by simple partition detection through less important edges or by an improved Markov clustering algorithm which uses the mathematical bootstrapping procedure [3].

One of the recent methods entails a flow-based technique which can deal both with the prediction of protein function and protein modularity analysis. There are several algorithms which have been developed with this concept. One of them is the 'Majority' method which considers the interactions of its neighbors and adopts the three most frequent annotations [58]. An extension of the above method, 'Neighborhood', employs a search for all the proteins within a particular radius to identify overrepresented functional annotations [59]. The usage of edge weights through gene expression data was done by Karaoz *et al*. [60]. Similar kind of weighted interaction network was used following a 'guilt-by-association' principle, wherein the functional flow was created from the annotated protein to the unannotated ones, through simulation [61]. Such kind of simulation of biological or functional flows within the network can be used as an essential tool of modeling to explore the dynamic signal transduction systems [3]. Moreover,

network flow simulations can predict complex network behavior under a realistic variety of external stimuli. A very important algorithm called CASCADE helps to detect the dynamic flow simulation of modularity analyses. CASCADE utilises the concept of occurrence probability and models a unique clustering methodology encompassing the biological and topological influence of each protein on the other. Occurrence probability brings out the distribution of the number of interactions necessary to link a pair of instant proteins in the network at a given time point [62].

The methods for the generation and analyses of the networks discussed as of now would be more accurate with a benchmarking of the data. Clustering techniques described here are based solely upon the graph theoretical properties without any real supervised data, thereby confirming their authenticity. However, *a priori* knowledge from amino acid and genomic sequences, protein structures and evolutionary profiles, gene expression and ontology annotation could be integrated with the PPI data to add to the analyses. Information about protein domains and localization has been used to successfully predict protein functions [63, 64]. A variety of high throughput data including microarray and protein complex data have been integrated to construct Bayesian models [65, 66], and Kernel based matrices have also been proposed [67, 68].

It is worth mentioning at this point that different clustering techniques and even the same technique with different parameters end up in giving disparate outcomes. Thus, validation of these clustering techniques is mandatory. Indeed, different clustering algorithms have been evaluated by several researchers in order to understand their potential to infer protein clusters from protein interaction networks [69, 70]. Jiang *et al.* [71], Zhang [3] have suggested different approaches to validate clustering methods, including validation based on agreement with annotated protein function databases, definition of clustering, the reliability of clusters, topological properties and the p-value from the hypergeometric distribution.

The visualization

There are several software and plug-ins, added therein, for the visualization of the real networks constructed in the form of graphs of the interconnected proteins. Researchers across the world have used Cytoscape (latest version 2.8.2) [72] and Gephi (latest version beta 0.8.2) [73]. Cytoscape has the plug-in, NETWORK ANALYZER [74], to compute values for the classical network centrality parameters like DC, CC and BC besides clustering coefficient, average path length (APL) and network diameter. Another important centrality measure, the EC, can be calculated via Gephi. The Java plug-in, cytoHubba [75], can be used to categorise the top ranked proteins/hubs in the network. Combined scores, from different parameters considered in the databases like STRING, can be taken as edge weights for computing Cytohubba scores. Several topological algorithms, viz. Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC), and Density of Maximum Neighborhood Component (DMNC), can be used to find the important hub proteins of the networks. To obtain the clusters of proteins after the network decomposition, the Molecular Complex Detection (MCODE) algorithm can be implemented to find the densely connected regions in the networks [52].

Analyses of pathogenic PINs: Intra- and Interspecies scenario

In order to gain insight into the infection strategies of pathogens, several intra-pathogenic and host-pathogen protein interaction networks have been generated and analysed over the last decade. This section will delineate the scenario of protein interaction network analysis of some of these species including viruses, bacteria and protozoan parasites (Table 1). Amongst these, the topology of intra-viral networks of different members of herpesvirus family (*viz.,* Kaposi's

sarcoma-associated herpesvirus (KSHV), Varicella-zoster virus (VZV), Epstein-Barr virus (EBV)), SARS-coronavirus (SARS-CoV), Hepatitis C virus (HCV) and Influenza A virus (H1N1 and H3N2) have been investigated by evaluating different network parameters like degree, APL, clustering coefficient and network diameter [76-82]. The analysis revealed that viral networks appear as single, highly coupled modules with relatively many hubs and few 'peripheral' nodes, in contrast to scale-free cellular networks having well-separated functional modules. This distinguishing network topology, may be essential for the formation of compact virions and functional viral complexes. However, it is unclear whether the disparity between viral and cellular network topology is a consequence of biological differences or the artifacts of experimental biases and errors.

Furthermore, the comparison of interactomes can lead to the identification of highly conserved interactions, critical for pathogenesis and thus, could serve as promising broad spectrum drug targets. For example, the comparison of intra-viral networks for herpesviruses enabled to identify a core set of highly conserved interactions, which mediate budding of capsids at the inner nuclear membrane of the host, and thus, could be promising targets for alternative herpesviral therapies.

The first large-scale intra-bacterial PPI networks were constructed and analyzed for *Helicobacter pylori* and subsequently, for several other bacterial pathogens, such as *Campylobacter jejuni*, *Treponema pallidum*, *Mycoplasma pneumonia, Mycobacterium tuberculosis*, and *Staphylococcus aureus* [44, 83-87]. The topological parameters (degree and BC) of undirected intra-bacterial networks, studied so far, revealed that bacterial networks are scale-free in nature, following a power law distribution. The evaluation of average clustering coefficient of bacterial protein networks (*eg*., *C.jejuni, M. tuberculosis etc.*) indicated that networks comprised of many clusters ie., subnetworks of highly interconnected proteins and comparative network analysis (CNA)

revealed that many of the subnetworks were conserved across organisms, identified using the NetworkBlast algorithm [84, 87]. It is worth to be noted that the proteins enriched in conserved subnetworks carry out specific Gene Ontology (GO) functions representing crucial functional pathways or protein complexes. Indeed, in *C. jejuni,* clustering of the conserved subnetworks using k-means algorithm followed by UPGMA identified core proteins having distinct cellular function. These core proteins were found to present in many subnetworks [84]. Thus, the organism's interaction network can be used to predict the function of the unannotated proteins or to map protein complexes and pathways involved in virulence, providing the directions for uncovering new drug targets [44, 84, 87]. Network topology was exploited to identify essential genes/proteins, which are crucial for replication, growth and viability of an organism, in different pathogenic species, including *S. aureus, C.jejuni, M. tuberculosis*, *Mycobacterium abscessus,* and various food and waterborne pathogens [44, 84, 88-90]. Proteins, encoded by essential genes, are hub proteins with many number of interactions in a network, and are also important for network integrity and stability, thus could be potential to be therapeutic targets.

The protein interaction networks for the above mentioned viruses, including *Human Immunodeficiency Virus* (HIV)-1 [76, 77, 81, 91-93] and different bacterial pathogens, such as *Bacillus anthracis*, *Francisella tularensis*, and *Yersiniapestis,* and *M.tuberculosis* with their human host have been studied [94-96]. The network topology analyses of host-pathogen systems indicated that both viral and bacterial proteins target human proteins which own higher degree and higher BC in the human protein interaction network. Viruses and bacteria both follow a common infection strategy of preferentially attacking hub and bottleneck proteins to impede host's essential biology [82]. Viruses tend to interact with host proteins which have higher degree and BC values compared to their bacterial counterparts. Identification of conserved

subnetworks in human-pathogen PPI [94] and GO functional analysis of pathogen-targeted human proteins, delineated a perfect picture of their infection strategy. Bacteria upholds infection in humans by foraying proteins involved in immune response thereby shattering human defense mechanism, whereas viruses exploit host's transcriptional machinery to propagate themselves within the host. It is worthy to mention at this point that most of the pathogen-targeted host proteins are those that play critical role in regulation of metabolic processes, such as cell-cycle regulation, nuclear transport and most importantly immune response.

Plasmodium falciparum, the causative agent of malaria in human, is the only protozoan parasite whose protein interaction network has been studied extensively [97-100]. Each study mostly focused on the identification and isolation of critical protein clusters/subnetworks or pathways, and also assigning the function of uncharacterized proteins. The study identified a group of proteins, such as chaperones, transcription factors and new surface proteins which are crucial for parasite's invasion and survival. Most of the proteins in the highly interconnected subnetworks were found to be involved in pathogenesis, perhaps the result of gene duplication event for maintaining its parasitic way of life. The identification of subnetworks was mainly done by using clustering coefficient, Markov clustering algorithm [97], clique percolation algorithm [98] and k-means clustering. It is worth mentioning at this point that plasmodium network stands distinct from other eukaryotic network because of its 'assortative' nature and bearing very less overlap with their interactomes. A very recent study [100] aimed at identifying important interacting proteins (IIPs) in Plasmodium PPI network, using various node centrality indices (degree, closeness, radiality, betweenness, eccentricity, stress, weinner index, centroid, assortativity and clustering coefficient) and network centrality indices (average distance, connectivity distribution, diameter and average clustering coefficient), followed by *in silico* knock-out analysis. The

highly interacting hub and central proteins, which are vital for network integrity as well as crucial for organism's survival, were considered as important proteins. These IIPs also play a vital role in stage specificity and were found to interact with several human proteins associated with multiple metabolic pathways, signaling pathways and infection mechanism. It has to be noted that human proteins targeted by pathogen are hubs in the human interactome and malfunctioning of the crucial host pathways results in clinical manifestation of malaria, which pose the interacting pathogen proteins as potential drug targets.

Applications in systems biomedicine

With the main target of the present review being the application of the PPI networks in biomedicine, cases to unravel the molecular basis of disease, by studying disease related subnetworks, have been reported. For instance, a new dimension has been given by such PPI network analysis to bring out the relationship of the pathogenic bacterium *Heliobacter pylori* with gastric carcinoma [101]. This has achieved a level of acceptance from the World Health Organization (WHO) and the International Agency for Research on Cancer consensus groups who have classified *H. pylori* as a definite biological carcinogen. The authors analysed the networks built upon the selected translated proteins of the expressed genes from databases and literatures. Their analyses reflected connectors of oncogenic proteins as hub and bottleneck proteins, mostly related to immune response governing the cell cycle, cell maintenance and proliferation, and transcription regulators [101].

An indirect study on a smaller scale upon *Salmoenlla* Pathogenecity Island Type III secretion system was carried out to build a methodology of targeting the indispensable proteins from amongst a conglomerate [42]. The authors constructed the network from the available

interactions from STRING database and analysed it with the common and rarely used centrality measures to decide upon the most indispensable one. They benchmarked their theoretical finding through analyses of networks built from the expressed gene products of two different microarray data and arrived at the same point with respect to such indispensable protein issue [42]. The outcome of these two works clearly would be the positive side of the analyses of PPI networks for generating systems biomedicine where the goal would be to harm the pathogen without harming the host and avoiding rapid development of antimicrobial resistance. The discussion on such issues takes us to a point wherein workers in this field would like to keep in mind few points while carrying out the related research. As indicated earlier, Lahiri *et al*. [42] have delineated a key methodology which could be followed with modifications as and when needed. A network constructed from a source has to be checked in for scale-freeness. The network can then be pruned to a core of proteins and/or top rankers from different centrality measures can be compared to unanimity. The finding therefrom can be benchmarked by other experimental omics data to corroborate. A selection of centrality measures would depend upon the requirement of the work. Following just some network analyses and trying to get a positive outcome, however, would abrogate the essentiality of PIN analyses.

Dependability of analyzed PINs

While there can be claims about the necessity of PIN analyses, a very important point needs to be considered to facilitate such claims. It is to be understood that the correctness of the analyses of such PIN would depend upon the correct construction of the network. Many such networks are being built based upon laboratory experimentation like yeast two-hybrid and mass spectrometry data generation. Moreover, networks built from various sources have extremely low overlap of

different high-throughput data generating manually curated databases. The other possibilities causing such error, down the line of network analyses, could be low reliability of literature curation and difficulties arising due to improper gene annotation and webpage data extraction [102]. As the above methods can be highly error prone, the dependability of PIN analyses become low [103]. In fact, there can be falsely reported interactions as well as left out interactions not being reported. Alarming false discovery rates (FDRs) of 10-20% and false negative rates (FNRs) of upto 50% are reported for yeast, worm and fly screens [104, 105]. However, such falsification of interactions could also crop from the low coverage of different comparative methods having noises leading to misinterpretation and erroneous integration of data [106]. An interesting concept on such comparative methods of interactions is ~~(to note)~~ that a comparison of the individual proteins interaction reveals a common tendency between methods manifested as global properties of the PINs [107]. To reduce such uncertainties of PIN construction from experimental data, two models have been proposed. These are the spoke and the matrix model of studying the bait connecting the prey. The former, connecting the bait along with the hit proteins, yields less false positives and is three times more accurate than the matrix model which connects all proteins. However, the latter yields more true positives as well [108]. A list of such sources of PIN can be obtained from literatures [102, 109].

Irrespective of the network construction, the analyses, however, can be of potential in cases of assessing the efficacy of a drug target, where a specific pathway is targeted to inhibit it. In this case, a perturbation of a dynamic network by inhibiting a specific pathway is manifested as a diversion to alternate pathways, as discussed in CASCADE [3]. However, the shortest path distance between important proteins remains the same and thus, proteins connecting other important ones in the network and thereby bridging them, have high BC. Instances of reduction

of such alternate pathways, keeping the core pathway intact as the shortest path, in metabolic network of *Mycobacterium leprae*, have been reported [110].

Concluding remarks

The present review has delineated a broad overview of analysing protein interaction networks of infectious diseases caused by viruses, bacteria and protozoan parasites. It entails the different methodologies which can be adopted by researchers while trying to analyse such networks. A thorough look of the review shows that most of the researchers resorted to only a handful of the techniques to conclude about important protein identification, pathway detection and functional prediction. It is imperative, however, that a benchmarking of these computationally predicted and analysed results would be mandatory for a better future towards non-conventional health intervention processes. For instance, Lahiri *et al*. [42] adopted several of these techniques and then validated with some biologically relevant high throughput microarray data. Days are not far when it would be a practice for the researchers to spread themselves and come up with new health intervention strategies to generate more accurate systems-based biomedicines.

Acknowledgement

Authors AP, AR and BS are thankful to Centre for Bioinformatics, Pondicherry University, Pondicherry, India and CL to the Department of Biological Sciences, Sunway University, Selangor, Malaysia for proving infrastructure facilities to carry on the work. AR and BS are indebted to Pondicherry University for their pre-doctoral fellowship.

References

1. Aderem A, Adkins JN, Ansong C, *et al*. A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. *MBio* 2011;**2**:e00325-10.
2. Maragakis LL, Perl TM. Acinetobacter baumannii: epidemiology, antimicrobial resistance, and treatment options. *Clin Infect Dis* 2008;**46**:1254-63.
3. Zhang A. Protein Interaction Networks: Computational Analysis. New York: Cambridge University Press, 2009.
4. Uetz P, Giot L, Cagney G, *et al*. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 2000;**403**:623-7.
5. Ho Y, Gruhler A, Heilbut A, *et al*. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 2002;**415**:180-3.
6. Ge H., UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions. *Nucleic Acids Res* 2000;**28**:e3.
7. Marcotte EM, Pellegrini M, Ng HL, *et al*. Detecting protein function and protein-protein interactions from genome sequences. *Science* 1999;**285**:751-3.
8. Overbeek R, Fonstein M, D'Souza M, *et al*. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 1999;**96**:2896-901.
9. Matthews LR, Vaglio P, Reboul J, *et al*. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or “Interologs”. *Genome Res* 2001;**11**:2120-6.
10. Wojcik J, Schächter V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 2001;**17**:S296-S305.

11. Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci USA* 2002;**99**:5896-901.

12. Smith GR, Sternberg MJ. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol* 2002;**12**:28-35.

13. Bock JR, Gough DA. Predicting protein-protein interactions from primary structure. *Bioinformatics* 2001;**17**:455-60.

14. Jansen R, Yu H, Greenbaum D, *et al*. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003;**302**:449-53.

15. Prasad TSK, Goel R, Kandasamy K, *et al*. Human protein reference database- update 2009. *Nucleic Acids Res* 2009;**37**:D767-72.

16. Persico M, Ceol A, Gavrila C, *et al*. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. BMC *Bioinformatics* 2005;**6**:S21.

17. Licata L, Briganti L, Peluso D, *et al*. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012;**40**:D857-61.

18. Stark C, Breitkreutz BJ, Reguly T, *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**:D535-9.

19. Salwinski L, Miller CS, Smith AJ, *et al*. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004;**32**:D449–51.

20. Pagel P, Kovac S, Oesterheld M, *et al*. The MIPS mammalian protein-protein interaction database. *Bioinformatics* 2005;**21**:832-4.

21. Alfarano C, Andrade CE, Anthony K, *et al*. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* 2005;**33**:D418-24.

22. Franceschini A, Szklarczyk D, Frankild S, *et al*. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013; **41**:D808-15.

23. Kerrien S, Aranda B, Breuza L, *et al*. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 2012;**40**:D841-6.

24. Warde-Farley D, Donaldson SL, Comes O, *et al*. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 2010;**38**:W214-20.

25. Schmitt T, Ogris C, Sonnhammer EL. FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res* 2014;**42**:D380-8.

26. Kamburov A, Stelzl U, Lehrach H, *et al*. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res* 2013;**41**:D793-800.

27. Wattam AR, Abraham D, Dalay O, *et al*. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 2014;**42**:D581-91.

28. Rid R, Strasser W, Siegl D, *et al*. PRIMOS: an integrated database of reassessed protein-protein interactions providing web-based access to in silico validation of experimentally derived data. *Assay Drug Dev Technol* 2013;**11**:333-46.

29. Kumar R, Nanduri B. HPIDB-a unified resource for host-pathogen interactions. BMC *Bioinformatics* 2010;**11**:S16.

30. Urban M, Pant R, Raghunath A, *et al*. The Pathogen-Host Interactions database (PHI-base): additions and future developments. *Nucleic Acids Res* 2015;**43**:D645-55.

31. Winnenburg R, Baldwin TK, Urban M, *et al*. PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res* 2006;**34**:D459-64.

32. Calderone A, Licata L, Cesareni G. VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res* 2015;**43**:D588-92.

33. Guirimand T, Delmotte S, Navratil V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res* 2015;**43**:D583-7.

34. Erdos P, Renyi A. On the evolution of random graphs, Publications of Mathematical Institute of Hungarian Academy of Science, 1960;**5**:17-61.

35. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* 1998; **393**:440- 2.

36. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. *Nature* 2000;**406**:378-482.

37. Jeong H, Mason SP, Barabasi AL, *et al*. Lethality and centrality in protein networks. *Nature* 2001;**411**:41-2.

38. Mason O, Verwoerd M. Graph theory and networks in Biology. *IET Syst Biol* 2007;**1**:89-119.

39. Pavlopoulos GA, Hooper SD, Sifrim A, *et al*. A tool for exploring and clustering biological networks. *BMC Res Notes* 2011;**4**:384.

40. Ozgur A, Vu T, Erkan G, *et al*. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* 2008;**24**:277-85.

41. Koschützki D, Schreiber F. Comparison of centralities for biological networks. In: Proceedings of the German Conference on Bioinformatics (GCB), 2004;**53**:199-206.

42. Lahiri C, Pawar S, Sabarinathan R, *et al*. Interactome analyses of Salmonella pathogenicity islands reveal SicA indispensable for virulence. *J Theor Biol* 2014; **363**:188-97.

43. Bonacich P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* 1972;**2**:113-120.

44. Cherkasov A, Hsing M, Zoraghi R, et *al*. Mapping the protein interaction network in Methicillin-resistant Staphylococcus aureus, *J Proteome Res* 2011;**10**:1139-1150.

45. Dickerson J, Pinney J, Robertson D. The biological context of HIV-1 host interactions reveals subtle insights into a system hijack. *BMC Syst Biol* 2010;**4**:1752.

46. Hahn MW, Kern, AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* 2005;**22**:803-6.

47. Han J, Bertin N, Hao T, *et al*. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 2004;**430**:88-93.

48. Hwang W, Cho YR, Zhang A, *et al*. A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms Mol Biol* 2006;**1**:24.

49. Dartnell L, Simeonidis E, Hubank M, *et al*. Robustness of the p53 network and biological hackers. *FEBS Lett* 2005;**579**:3037-42.

50. Xu K, Bezakova I, Bunimovich L, *et al*. Path lengths in protein-protein interaction networks and biological complexity. *Proteomics* 2011;**11**:1857-67.

51. Asif W, Qureshi A, Iqbal M, *et al*. On the complexity of average path length for biological networks and patterns. *Int J Biomath* 2014; 7.

52. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003;**4**:2.

53. Seidman SB. Network structure and minimum degree. Social Networks 1983; 5:269-87.

54. Bollobas B. The evolution of sparse graphs. In Graph Theory and Combinatorics, Proceeding Cambridge Combinatorial Conference in honor of Paul Erdos, 1984, 35-57.

55. Enright AJ, van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;**30**:1575-84.

56. Domingues F, Rahnenfuhrer J, Lengauer T. Automated clustering of ensembles of alternative models in protein structure databases. *Protein Eng Des Sel* 2004;**17**:537-43.

57. Tornow S, Mewes HW. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res* 2003;**31**:6283-9.

58. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nat Biotechnol* 2000;**18**:1257-61.

59. Hishigaki H, Nakai K, Ono T, *et al*. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast* 2001;**18**:523-31.

60. Karaoz U, Murali TM, Letovsky S, *et al*. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci USA* 2004;**101**:2888-93.

61. Nabieva E, Jim K, Agarwal A. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 2005;**21**:i302-10.

62. Hwang W, Kim T, Ramanathan M, *et al*. Bridging centrality: graph mining from element level to group level. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery& Data Mining (KDD08), 2008;336-44.

63. Chen X, Liu M, Ward R. Protein function assignment through mining cross-species protein-protein interactions. *PLoS One* 2008;**3**:e1562.

64. Nariai N, Kasif S. Context specific protein function prediction. *Genome Inform* 2007; **18**:173-82.

65. Troyanskaya OG, Dolinski K, Owen AB, *et al*. A Bayesian framework for combining heterogeneous data sources for gene function prediction. *Proc Natl Acad Sci USA* 2003; **100**:8348-53.

66. Chen Y, Xu D. Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. *Nucleic Acids Res* 2004;**32**:6414-24.

67. Lanckriet GRG, Deng M, Cristianini N, *et al*. Kernel-based data fusion and its application to protein function prediction in yeast. Pacific Symposium on Biocomputing, 2004;9.

68. Tsuda K, Shin HJ, Schoelkopf B. Fast protein classification with multiple networks. 2005;**21**:ii59-65.

69. Brohee S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 2006;**7**:488.

70. Moschopoulos CN, Pavlopoulos GA, Iacucci E, *et al*. Which clustering algorithm is better for predicting protein complexes? *BMC Res Notes* 2011;**4**:549.

71. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A Survey. IEEE Transactions on Knowledge and Data Engineering (TKDE) 2004;**16**:1370-86.

72. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498-504.

73. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the International AAAI Conference on Weblogs and Social Media. San Jose, CA, North America (ICWSM09), 2009.

74. Assenov Y, Ramirez F, Schelhorn SE, *et al*. Computing topological parameters of biological networks. *Bioinformatics* 2008;**24**:282-4.

75. Lin CY, Chin CH, Wu HH, *et al*. Hubba: hub objects analyzer-a framework of interactome hubs identification for network biology. *Nucleic Acids Res* 2008;**36:**W438-43.

76. Uetz P, Dong YA, Zeretzke C, *et al*. Herpesviral protein networks and their interaction with the human proteome. *Science* 2005;**311**:239-42.

77. Calderwood MA, Venkatesan K, Xing L, *et al*. Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci USA* 2007;**104**:7606-11.

78. Bailer SM, Hass J. Connecting viral with cellular interactomes. *Curr Opin Microbiol* 2009;**12**:453-9.

79. von Brunn A, Teepe C, Simpson JC, *et al*. Analysis of intraviral protein-protein interactions of the SARS coronavirus ORFeome. *PLoS One* 2007;**2**:e459.

80. Meyniel-Schicklin L, de Chassey B, André P, *et al*. Viruses in interactomes in translation. *Mol Cell Proteomics* 2012;11.

81. Shapira SD, Gat-Viks I, Shum BO, *et al*. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* 2009;**139**:1255-67.

82. Tekir SD, Çakır T and Ülgen KÖ. Infection strategies of bacterial and viral pathogens through pathogen-human protein-protein interactions. *Front in Microbiol*, 2012, **3**:46,1-11.

83. Rain JC, Selig L, De Reus H, *et al*. The protein-protein interaction map of Helicobacter pylori. *Nature* 2001;**409**:211-5.

84. Parrish JR, Yu J, Liu G, *et al*. A proteome-wide protein interaction map for Campylobacter jejuni. *Genome Biol* 2007;**8**:R130.

85. Titz B, Rajagopala SV, Goll J, *et al*. The binary protein interactome of Treponema pallidum - the syphilis spirochete. *PLoS One* 2008;**3**:e2292.

86. Kühner S, van Noort V, Betts MJ, *et al*. Proteome organization in a genome-reduced bacterium. *Science* 2009;**326**:1235-40.

87. Wang Y, Cui T, Zhang C, *et al*. Global protein-protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv. *J Proteome Res* 2010;**9**:6665-77.

88. 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.

89. Shanmugham B, Pan A. Identification and characterization of potential therapeutic candidates in emerging human pathogen Mycobacterium abscessus: a novel hierarchical in silico approach. *PLoS One* 2013;**8**:e59126.

90. Jadhav A, Shanmugham B, Rajendiran A, *et al*. Unraveling novel broad-spectrum antibacterial targets in food and waterborne pathogens using comparative genomics and protein interaction network analysis. *Infect Genet Evol* 2014;**27**:300-8.

91. de Chassey B, Navratil V, Tafforeau L, *et al*. Hepatitis C virus infection protein network. *Mol Syst Biol* 2008;**4**:230.

92. Emamjomeh A, Goliaei B, Zahiriab J, *et al*. Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method. *Mol Biosyst* 2014;**10**:3147-54.

93. Bandyopadhyay S, Ray S, Mukhopadhyay A, *et al*. A review of in silico approaches for analysis and prediction of HIV-1-human protein-protein interactions. *Brief Bioinform* 2014.

94. Dyer MD, Neff C, Dufford M, *et al*. The human-bacterial pathogen protein interaction network of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. *PLos One* 2010; **5**:e12089.

95. Yang H, Ke Y, Wang J, *et al*. Insight into bacterial virulence mechanisms against host immune response via the Yersinia pestis-human protein-protein interaction network. *Infect Immun* 2011;**79**:4413-24.

96. Zhou H, Gao S, Nguyen NN, *et al*. Stringent homology-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. *Biol Direct* 2014;**9**:5.

97. Wuchty S, Ipsaro JJ. A Draft of Protein Interactions in the Malaria Parasite P. falciparum. *J Proteome Res* 2007;**6**:1461-70.

98. Wuchty S, Adams JH, Ferdig MT. A comprehensive Plasmodium falciparum protein interaction map reveals a distinct architecture of a core interactome. *Proteomics* 2009;**9**: 1841-9.

99. Wuchty S. Computational Prediction of Host-Parasite Protein Interactions between P. falciparum and H. sapiens *PLoS One* 2011;**6**:e26960.

100. Bhattacharyya M, Chakrabarti S. Identification of important interacting proteins (IIPs) in Plasmodium falciparum using large-scale interaction network analysis and in silico knock-out studies. *Malaria Journal* 2015;**14**:70.

101. Kim KK, Kim HB. Protein interaction network related to *Helicobacter pylori* infection response. *World J Gastroenterol* 2009;**15**:4518-28.

102. Cusick ME, Yu H, Smolyar A, *et al*. Literature-curated protein interaction datasets. *Nat Methods* 2009;**6**:39-46.

103. Rinner O, Mueller LN, Hubálek M, *et al*. An integrated mass spectrometric and computational framework for the analysis of protein interaction networks. *Nat Biotechnol* 2007;**25**:345-52.

104. Huang H, Bader JS. Precision and recall estimates for two-hybrid screens. *Bioinformatics* 2009;**25**:372-8.

105. Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? *Genome Biol* 2006;**7**:120.

106. Gentleman R, Huber W. Making the most of high-throughput protein-interaction data. *Genome Biol* 2007;**8**:112.

107. Hoffmann R, Valencia A. Protein interaction: same network, different hubs. *Trends Genet* 2003;**19**:681-3.

108. Bader GD, Hogue CW. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* 2002;**20**:991-7.

109. Sanderson CM. The cartographers toolbox: building bigger and better human protein interaction networks. *Brief Funct Genomic and Proteomic* 2009;**8**:1-11.

110. Verkhedkar KD, Raman K, Chandra NR, *et al*. Metabolome Based Reaction Graphs of M. tuberculosis and M. leprae: A Comparative Network Analysis. *PLoS One* 2007;**2**:e881.

Table legend

Table1:Mode of protein interaction network analysis in intra-pathogenic and host-pathogen systems

Supplementary Table1: The number of intra-pathogenic and host-pathogen PPIs from different resources along with pathogens' genome size

Figure legend

Figure 1: Schematic representation summarizing different computational approaches to analyze protein interaction networks.

Table 1: Mode of protein interaction network analysis in intra-pathogenic and host-pathogen systems

Pathogen	Disease	Mode of Network analysis	References
Kaposi's sarcoma-associated herpesvirus (KSHV) [φ,Ψ]	Kaposi sarcoma, B-cell lymphomas	[φ]DC,BC,DD,APL,CCf,ND [Ψ] DC	[76,80]
Varicella zoster virus (VZV) [φ, Ψ]	Chickenpox, shingles	[φ]DC,BC,DD,APL,CCf,ND [Ψ] DC	[76,80]
Epstein-Barr virus (EBV) [φ,Ψ]	Mononucleosis	[φ]DC,BC,DD,APL,CCf,ND [Ψ] DC,APL,CCf	[77,80]
Severe acute respiratory syndrome-coronavirus (SARS-CoV)[φ]	Severe acute respiratory syndrome	[φ]DC,BC,DD,APL,CCf,ND	[79,80]
Hepatitis C virus (HCV)[φ,Ψ]	Hepatitis	[φ]DC,BC,DD,APL,CCf,ND [Ψ] DD,BC,CC,APL, CCf,NC,Str,Ecc,Ra, PA(KEGG),GO,DA	[80,91,92]
Influenza A virus (H1N1, H3N2) [φ,Ψ]	Influenza	[φ]DC,BC,DD,APL,CCf,ND [Ψ] SNI,PA,GO	[80,81]
Human Immunodeficiency Virus (HIV)-1[Ψ]	Acquired Immunodeficiency Syndrome (AIDS)	[Ψ] DC,BC,DD,SNI,GO, PA(KEGG)	[93]
Helicobacter pylori [φ]	Gastritis, peptic ulcer and gastric cancer	[φ] CNA,DA	[83]
Campylobacter jejuni [φ]	Gastroenteritis	[φ]DC,CCf,SNI,CNA, GO,EPI	[84]
Treponema pallidum [φ]	Syphilis	[φ]SNI,CNA	[85]
Mycoplasma pneumoniae [φ]	Atypical pneumonia	[φ]SNI	[86]
Mycobacterium tuberculosis [φ,Ψ]	Tuberculosis	[φ]DC,CC,DD,APL,CCf, ND,Str,SNI,CNA [Ψ] DC,GO,PA(IntAct)	[87,96]

Table 1 continued

Pathogen	Disease	Mode of Network analysis	References
Staphylococcus aureus[φ]	Abscesses, Furuncles, Atopic dermatitis	[φ]DC,BC,EPI	[44]
Bacillus anthracis [Ψ]	Anthrax	[Ψ]DC,BC,CNA,GO	[94]
Francisella tularensis[Ψ]	Pneumonia	[Ψ]DC,BC,CNA,GO	[94]
Yersinia pestis[Ψ]	Pneumonic, septicemic, and bubonic plagues	[Ψ]DC,BC,APL,CNA,GO, PA	[94,95]
Plasmodium falciparum[φ,Ψ]	Malaria	[φ]DC,BC,CC,APL,CCf, Str,Ecc,Ra,Ass,WI , Cn,ND,GO,PA [Ψ] SNI, GO	[99,100]

[φ] - analyses of intra-pathogenic system, [Ψ] - analyses of host-pathogen systems

DC-degree centrality, BC-betweenness centrality, DD-degree distribution, APL-average path length, CCf-clustering coefficient, CC-closeness centrality, ND-network diameter, NC-neighbor connectivity, Str-stress, Ecc-eccentricity, Ra-radiality, Ass-assortativity, WI-Weiner Index, Cn-centroid, DA-domain analysis; CNA-comparative network analysis, PA-pathway analysis, SNI-subnetwork identification, EPI-essential protein identification, GO-gene ontology analysis.

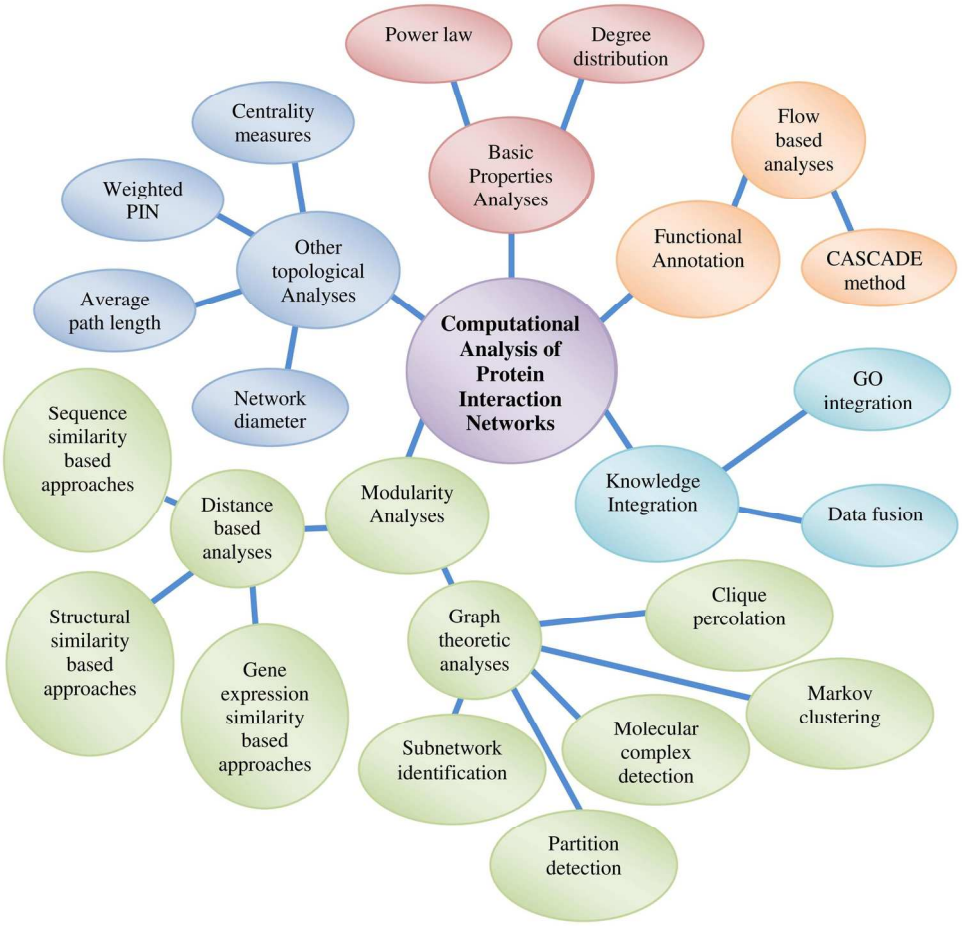

Figure 1: Schematic representation summarizing different computational approaches to analyze protein interaction networks
169x158mm (300 x 300 DPI)

Supplementary Table1: The number of intra-pathogenic and host-pathogen PPIs from different resources along with pathogens' genome size

Pathogen	Genome Size	Interactions		Resources
Bacillus anthracis	5.2 Mb	3036$^{\varphi}$	3020$^{\psi}$	PATRIC
Campylobacter jejuni NCTC 11168	1.6 Mb	24023$^{\varphi}$		PATRIC
Clostridum botulinum	3.9 Mb	10$^{\varphi}$	8$^{\psi}$	PATRIC
Escherichia coli O157:H7	5.6 Mb	3027$^{\varphi}$	12$^{\psi}$	PATRIC
Francisella tularensis	1.9 Mb	1312$^{\varphi}$	1312$^{\psi}$	PATRIC
Helicobacter pylori	1.7 Mb	2784$^{\varphi}$	4$^{\psi}$	PATRIC
Listeria monocytogenes	2.9 Mb	5$^{\varphi}$	5$^{\psi}$	PATRIC
Mycobacterium tuberculosis	4.4 Mb	8042$^{\varphi}$		[82]
Mycoplasma pneumonia	0.81 Mb	178$^{\varphi}$		[82]
Pasteurella multocida	2.4 Mb	12$^{\psi}$		HPIDB
Salmonella enterica Typhi	4.8Mb	10$^{\varphi}$	5$^{\psi}$	PATRIC
Shigella flexneri	4.6 Mb	191$^{\varphi}$	41$^{\psi}$	PATRIC
Staphylococcus aureus	2.9 Mb	16$^{\varphi}$	21$^{\psi}$	PATRIC
Streptococcus pneumoniae TIGR4	2.1 Mb	429$^{\varphi}$	12$^{\psi}$	PATRIC
Treponema pallidum	1.1 Mb	3649$^{\varphi}$		[82]
Vibrio cholera O1 biovar El Tor N16961	4 Mb	9$^{\varphi}$	1$^{\psi}$	PATRIC
Yersinia pestis	4.7 Mb	3948$^{\varphi}$	4018$^{\psi}$	PATRIC, HPIDB
Epstein–Barr virus	171.8 Kb	220$^{\varphi}$		BioGRID
Hepatitis C virus	9.6 Kb	111$^{\varphi}$		BioGRID
Human immunodeficiency virus	9.1 Kb	1195$^{\varphi}$		BioGRID
Influenza A virus (H1N1)	13.6 Kb		4067$^{\psi}$	HPIDB
Kaposi's sarcoma-associated herpesvirus	138 Kb	142$^{\varphi}$		BioGRID
Severe acute respiratory syndrome-coronavirus	29.8 Kb	65$^{\varphi}$		[82]
Varicella zoster virus	125Kb	173$^{\varphi}$		[82]
Plasmodium falciparum	22.9 Mb	4750$^{\varphi}$	367$^{\psi}$	[100]

φ and ψ represent the number of interactions for intra-pathogenic and host-pathogen systems, respectively

Supplementary Table1: The number of intra-pathogenic and host-pathogen PPIs from different resources along with pathogens' genome size

Pathogen	Genome Size	Interactions		Resources
Bacillus anthracis	5.2 Mb	3,036$^{\varphi}$	3020$^{\psi}$	PATRIC
Campylobacter jejuni NCTC 11168	1.6 Mb	24,023$^{\varphi}$		PATRIC
Clostridum botulinum	3.9 Mb	10$^{\varphi}$	8$^{\psi}$	PATRIC
Escherichia coli O157:H7	5.6 Mb	3,027 $^{\varphi}$	12$^{\psi}$	PATRIC
Francisella tularensis	1.9 Mb	1,312 $^{\varphi}$	1312$^{\psi}$	PATRIC
Helicobacter pylori	1.7 Mb	2,784$^{\varphi}$	4$^{\psi}$	PATRIC
Listeria monocytogenes	2.9 Mb	5$^{\varphi}$	5$^{\psi}$	PATRIC
Mycobacterium tuberculosis	4.4 Mb	8,042$^{\varphi}$		[82]
Mycoplasma pneumonia	0.81 Mb	178$^{\varphi}$		[82]
Pasteurella multocida	2.4 Mb	12$^{\psi}$		HPIDB
Salmonella enterica Typhi	4.8Mb	10$^{\varphi}$	5$^{\psi}$	PATRIC
Shigella flexneri	4.6 Mb	191$^{\varphi}$	41$^{\psi}$	PATRIC
Staphylococcus aureus	2.9 Mb	16$^{\varphi}$	21$^{\psi}$	PATRIC
Streptococcus pneumoniae TIGR4	2.1 Mb	429$^{\varphi}$	12$^{\psi}$	PATRIC
Treponema pallidum	1.1 Mb	3,649$^{\varphi}$		[82]
Vibrio cholera O1 biovar El Tor N16961	4 Mb	9$^{\varphi}$	1$^{\psi}$	PATRIC
Yersinia pestis	4.7 Mb	3,948$^{\varphi}$	4018$^{\psi}$	PATRIC, HPIDB
Epstein–Barr virus	171.8 Kb	220$^{\varphi}$		BioGRID
Hepatitis C virus	9.6 Kb	111$^{\varphi}$		BioGRID
Human immunodeficiency virus	9.1 Kb	1,195$^{\varphi}$		BioGRID
Influenza A virus (H1N1)	13.6 Kb		4067$^{\psi}$	HPIDB
Kaposi's sarcoma-associated herpesvirus	138 Kb	142$^{\varphi}$		BioGRID
Severe acute respiratory syndrome-coronavirus	29.8 Kb	65$^{\varphi}$		[82]
Varicella zoster virus	125Kb	173$^{\varphi}$		[82]
Plasmodium falciparum	22.9 Mb	4,750$^{\varphi}$	367$^{\psi}$	[100]

φ and ψ represent the number of interactions for intra-pathogenic and host-pathogen systems, respectively