A review of in silico approaches for analysis and prediction of HIV-1-human protein–protein interactions

Sanghamitra Bandyopadhyay*, Sumanta Ray*, Anirban Mukhopadhyay and Ujjwal Maulik

Abstract

The computational or in silico approaches for analysing the HIV-1-human protein–protein interaction (PPI) network, predicting different host cellular factors and PPIs and discovering several pathways are gaining popularity in the field of HIV research. Although there exist quite a few studies in this regard, no previous effort has been made to review these works in a comprehensive manner. Here we review the computational approaches that are devoted to the analysis and prediction of HIV-1-human PPIs. We have broadly categorized these studies into two fields: computational analysis of HIV-1-human PPI network and prediction of novel PPIs. We have also presented a comparative assessment of these studies and proposed some methodologies for discussing the implication of their results. We have also reviewed different computational techniques for predicting HIV-1-human PPIs and provided a comparative study of their applicability. We believe that our effort will provide helpful insights to the HIV research community.

Key words: HIV-1-human PPI Network; computational PPI prediction; HIV dependency factor; association rule mining; biclustering; random forest classifier; semi supervised classification; rank aggregation; topological properties of network

Introduction

Since 1980s, rapid progression has been noted in the field of HIV/AIDS research. HIV belongs to a special class of viruses called retroviruses. Within this class, HIV is placed in the sub-group of lentiviruses. HIV-1 virus contains a single-stranded RNA genome, which codes for primarily 19 proteins; thus, it mostly relies on human cellular functions for most activities during its life cycle. The RNA genome, consisting of seven structural landmarks (LTR, TAR, RRE, PE, SLIP, CRS and INS) and nine genes (gag, pol, env, tat, rev, nef, vif, vpr and vpu), encodes 19 proteins. Among these genes, gag, pol and env contain information needed to make the structural proteins for new virus particles [1].

One of the important areas of HIV research is the detailed understanding of HIV biology, its evolution and origins [1–6] and...
the process of replication in host cell. Like all other viruses, HIV-1 must hijack the host cellular machinery for increasing the production of viral genomic material. Initially, HIV-1 virus attaches the envelop protein gp120, which is embedded in the outer membrane of the HIV virion, to the human receptor protein CD4 T+ cells. To pass through the cell membrane, binding to a second receptor is also required. Two different coreceptors are also involved in the binding process, namely, CCR5, a chemokine receptor that serves as a coreceptor early in the infection and chemokine receptor (CXCR4) that serves as a coreceptor later. After entering the cell, an enzyme called Reverse Transcriptase is responsible for initiating the reverse transcription of a single stranded viral RNA to produce a double stranded viral DNA. After that, the viral DNA enters into the CD4 cell nucleus and integrates into the cell DNA. An enzyme called integrase acts as a key regulator in this process. After that, long chain of HIV proteins is created through transcription and translation by using the machinery of the host CD4 cell. An HIV enzyme called protease cuts up the long chain to produce smaller HIV proteins, which combine with HIV RNA to form a new virus particle. Viral proteins interact with human proteins at every vital stage of the HIV life cycle and form a complex network of molecular events, including virus–host protein–protein interactions (PPIs) [7, 8]. This intricate network of host and viral proteins are thus considered as an essential source for understanding the inherent mechanism of viral replication cycle [9].

In silico approaches have been developed as an alternative for conducting research at the system level. Most of the approaches primarily focus on the analysis of HIV-1-human PPI network, which is constructed from the available existing information catalogued in the HIV-1-Human Protein Interaction Database (HHPID) [10], published in 2009. There are several approaches that use this database to uncover various relationships between the viral proteins, between the virus and the host and ultimately between the various host proteins by exploring diverse methodologies. This valuable information remains in different published literature in an incoherent way. For an individual researcher, it is difficult to efficiently access this information and also time-consuming to acquire an overview in this area. Thus, it would be beneficial to collect all the works and compare different approaches in a single article.

Here we provide a comprehensive survey of various computational approaches that mainly focus on this HIV-1-human PPI network. To retain the computational flavour we restrict our analysis at the systems level. Despite being overwhelmed with dozens of experimental and computational results, our aim is to provide a complete view of different works on computational analysis of the HIV-1-human PPI network, constructed from HHPID data set. We broadly divide all these approaches in two categories: network analysis at systems level and prediction of novel PPIs between HIV-1 and human proteins using various computational approaches. Figure 1 shows the classification of the system-level approaches in these two categories and provides a simple overview of all the approaches with year of publication. Since 2006, a number of researchers have provided their valuable opinions and shed light on different issues for uncovering the inherent relationship between HIV-1 and human proteins. Here we try to review most of these works and accumulate their results to give a simplified view.

### The HIV-1-human protein interaction network

The National Institute of Allergy and Infectious Diseases has published the ‘HIV1-Human Protein Interaction Database’ (HHPID) [10] in 2009. This database provides a concise but detailed summary of all known interactions including their brief descriptions. It is a valuable resource for the HIV-1 research community. The database was made publicly available at NCBI’s website, http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/ in 2009. The HHPID database provides comprehensive data consisting of 5127 interactions between 19 HIV-1 proteins and 1432 human proteins (January, 2013). This curated database was constructed from approximately 3200 papers published between 1984 and 2007 and was validated by being linked to several publications. Over 14 312 PMID references to the original articles were collected for validating all the interactions in this growing database [10]. Moreover, each interaction is associated with a specific interaction type like ‘upregulate’, ‘activate’, ‘bind’, ‘inhibited by’ etc. A visualization of this database can be seen in Figure 2. We have extensively searched and found 68 unique interaction types among the 5127 interactions. The bipartite network shown in Figure 2a is constructed from this data set and consists of 19 red nodes, which represent HIV-1 proteins, and 1432 green nodes, which represent human proteins. We broadly divide all the interaction types into regulatory classes: regulating, regulated by and bidirectional (regulation is in both ways), where the three classes contain 34, 25 and 9 interaction types, respectively. We create three subnetworks, each corresponding to an interaction class. Figure 2b–d) show interaction types of the three classes, respectively. The degree distribution

---

**Figure 1.** Categorization of system-level approaches in the analysis of HIV-1-human PPI network.
in each subnetwork approximately obeys power law i.e. the number of nodes with large degree values are less frequent, whereas the number of nodes with lower degree values are more frequent. Following the distribution of interaction types in each subnetwork, 'upregulates', 'inhibits', 'downregulates', 'activates', 'binds', 'interacts with', 'inhibited by' and 'processed by' seem to be more common interactions in HHPI. Additionally, a significant number of interactions are observed to be associated with class-1 regulation type than the other two types of regulations.

To understand which of the subnetworks mimics the topological properties of the whole HIV-1-human bipartite network, we plot relative degree of HIV-1 proteins in each subnetwork against the whole bipartite network. This is shown in Figure 3. It is evident from the figure that there is a strong correlation between degree of subnetwork 1 or subnetwork 2 and the global
network \( (R^2 = 0.9949 \) and 0.9895 for subnetwork 1 and subnetwork 2, respectively). The fitted regression line \( (R^2 = 0.8104) \) for subnetwork 3, which is constructed by considering the class 3 type of interactions, asserts the existence of correlation in degree between subnetwork 3 and the whole bipartite network but not so closely as subnetwork 1 and subnetwork 2. This suggests that the relative degree of HIV-1 proteins retain the same pattern strongly in subnetwork 1 and in subnetwork 2. We have also studied degree and betweenness centrality distributions of HIV-1 interacting human proteins specific to each of these three subnetworks. For this purpose we collect the human PPI data set from Human Protein Reference Database (HPRD) [11], which consists of 39,240 interactions among 9589 human proteins. Figure 4a and b show, respectively, the degree and betweenness centrality distribution of HIV-1 interacting human proteins belonging to the three subnetworks in the corpus of entire human PPI network. The motivation behind this is to see to what extent the HIV-1 interacting human proteins, associated with specific type of regulation, are connected in the whole human PPI network. Not surprisingly, no such significant difference is observed in degree and betweenness centrality pattern. This suggests that the topological characteristics of HIV-1 interacting human proteins are weakly correlated with specific regulation type.

Figure 3. Relative degree of 19 HIV-1 proteins in three subnetworks versus relative degree of whole bipartite network.

Figure 4. Connectivity and betweenness centrality distributions of HIV interacting human proteins belonging to three subnetworks (positive, bidirected and negative) measured in the context of entire human PPI network. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.
A review on HIV-1-human PPI network analysis approaches

Here we have reviewed several pioneering works dedicated to the broad analysis of the HIV-1-human protein interaction network. Since 2009 with the introduction of HHPID data set, several works were published that use this data set from different perspectives. Almost all of them primarily focus on building a bipartite network constituting HIV-1 and human proteins as nodes and interactions between them as edges. The created network is analysed to gain a deeper insight into the HIV research that includes HIV-host perturbation, biological importance of the host proteins that show strong association with specific HIV interactions, different interaction patterns of host proteins with HIV-1 proteins and several other issues. The works we have reviewed in this section use the HHPIID data set as well as use different data sources that are integrated with HHPID.

Host factors essential for HIV replication

In Pinney et al. [9] an attempt was made to uncover different new developments in HIV research that provide a helpful insight into the viral perturbation in the host proteome. For this purpose, they have used the existing knowledge of HIV-host interactions documented in HHPID [10]. The HHPID data set is visualized here as a detailed map of viral perturbation of the host proteome in the context of gene ontological annotations of human proteins in this data set. The over-represented biological process terms of the human proteins in this data set reveal how the various parts of the viral replication cycle are delegated to different genes and the way in which HIV-1 gene products co-operate to target specific parts of the human cellular system. A minimum spanning tree (MST) is constructed by calculating the semantic distance [12] between each pair of HIV-1 interacting gene products. As a consequence the gene products that are close to each other in the MST are more likely to be involved in common biological process.

Pinney et al. [9] found little overlap between HIV-dependency factor (HDF) (which is basically a class of human proteins that are essential for HIV replication) identified by Brass et al. [13], Konig et al. [14] and Zhou et al. [15] and HIV-1 interacting human proteins reported in HHPID (shown in Figure 5). This discrepancy between HDF data sets and HHPID data is not surprising because the experimental screening procedure for HDF detection mostly relies on the cell line, HIV strain used and particular experimental protocol. However, 131 human proteins are believed to interact with HIV-1, as supported by at least two studies; these have a better chance of being a host factor or an essential protein. Essential or lethal proteins are those, deletion of which will result in lethality or infertility, i.e. the organism cannot survive without these proteins [16, 17]. Lethality of proteins is linked with different unique topological characteristics of proteins in the human PPI network [18–21]. It is demonstrated that the proteins with high degree and high betweenness centrality have a tendency of being essential [22–27]. In general ‘centrality-lethality’ rule states that deletion of a hub protein is more likely to be lethal than deletion of a non-hub protein [24].

The underlying base of the ‘centrality-lethality’ rule is set up by some interactions between proteins that are indispensable for the survival or reproduction of an organism. For examining the involvement of essential proteins with the HIV-interacting ones, Dickerson et al. [28] study the relationship between protein essentiality and degree/betweenness centrality. For this, they have considered mouse genome knockout data from [29]. Here a human gene is considered as ‘essential’ if a knockout of its mouse ortholog has a lethal impact [30]. They surprisingly observed that the HIV-1 interacting proteins turn out to be non-essential according to the definition of essentiality in [30]. The possible reason behind this may be that the essential proteins may follow poor correlation with degree and betweenness centrality.

It is interesting to see whether the HIV-1 interacting proteins play important roles in human biological processes or have certain functionalities that make them essential. For this, we have adopted a rank aggregation scheme to rank all the proteins that are predicted to interact with HIV-1, and are supported by at least two of the four studies (shown in Figure 5). We have used five topological metrics, viz., density, betweenness centrality, clustering coefficient, eigenvector centrality and average density of neighbour nodes, to rank all the proteins. A weighted rank aggregation technique [31] is used to build an ordered list of proteins. Figure 6 shows the result of applying the rank aggregation algorithm. It appears from the figure that most of the proteins that are predicted to interact with HIV-1 and are supported by at least three studies get better rank than the others.

Table 1 shows 11 proteins interacting with HIV-1 and are supported by at least three literature studies. Columns show the ranked list of these proteins corresponding to each metric. Four proteins, viz., ‘CD4’, ‘TCEB3’, ‘CHST1’ and ‘DDX3X’ got aggregated rank above 50. The conserved RNA helicase DDX3X has rank 106, which indicates that it is not so much essential as others in the list. DDX3X plays a major role in nuclear export of viral RNA transcripts into cellular cytoplasm. It is known to inhibit antiviral immune mechanisms, which shows its importance in viral replication as well as antiviral activity. Interestingly in [32] it is reported that knockdown of DDX3X drastically deteriorates HIV-1 replication in host cells. CD4 is a glycoprotein found on the surface of immune cells such as T helper cells and is essential for building up the human immune system. Although CD4 was placed in the 78th position in the optimal ranking, it plays a key role in transferring signal to other types of immune cells. Similarly, CXCR4, which acts as an important co-receptor for HIV-1 entrance in the host cell, got 35th position in the optimal ranking list. It is notable that CXCR4 appears in the first position in the ranking list corresponding to the degree or connectivity. All the four studies support the interaction with protein RELA.

We have performed a significance test to investigate whether the ranking of HIV-1 interacting proteins is significantly better than a random set of proteins. For this purpose,
we first rank all 129 HIV-1 interacting proteins supported by at least two studies in the corpus of all human proteins. Ranking is performed by considering the five topological metrics, used earlier. We also collected 129 human proteins randomly from the whole human protein set and performed ranking based on these five topological metrics. We have carried out this procedure 300 times and got 300 ranked lists of random proteins. These are then compared with the ranking of HIV-1 interacting proteins by using Wilcoxon Rank Sum test. The resulting P-value is significantly low (8.6733e-10), which indicates that the ranking of HIV-1 interacting proteins is significantly better than a random set of proteins.

Ranking of proteins apparently induces essentiality, possibly because of its affinity to get strongly correlated with unique topological characteristics. So, the HIV-1 interacting proteins, which got low ranks, are highly important for facilitating HIV-1 infection.

**Topological insights into HIV interacting host factors**

Besides the biological and ontological analysis of the HHPID data set, there are several works that use this catalogued information for the purpose of analysing the network to gain interesting topological insights. For example, Dickerson et al. [28] analysed the HIV-1-human bipartite network based on its topological properties and showed an association of interaction pattern of HIV-1 and human proteins. Their analysis mainly concentrated on the general properties of the HIV-1-human PPI interactions and explored the molecular specificity of the HIV-1-host perturbation. The analysis showed how the HIV-1 proteins have a tendency to interact with highly connected (‘hubs’) and central (‘bottlenecks’) proteins. The ‘hubs’ are proteins that have high degree and ‘bottlenecks’ are proteins with high betweenness centrality. The degree centrality represents the number of connected edges, i.e. the number of protein interactions measured per node. Betweenness centrality is used to measure the centrality of a node in the network by counting the number of shortest paths that go through that node. In other words, how many shortest paths would increase in length if the node is removed from the network [33]. In Pinney et al. [9], ten over-represented biological processes, viz., cell differentiation, intracellular signalling cascade, apoptosis, immune response, regulation of cell cycle, DNA packaging, inflammatory response, protein-DNA complex assembly, chemotaxis and nucleosome assembly in HIV-1 interactions were identified. A direct consequence was observed in [28] between degree and betweenness centrality of human proteins and the over-representation of those proteins in fundamental biological processes. The mean degree and betweenness centrality of human proteins associated with the above biological processes were $7.27 \times 10^5$ and $7.40 \times 10^5$, whereas these values reduced to $2.63 \times 10^5$ and $2.33 \times 10^5$ when whole human protein interaction network is considered. Therefore, it is obvious that the proteins involved in fundamental biological processes which are over-represented...
in HIV-1 interactions, are more connected and central. In Dijk et al. [34] associations between proteins involved in RNA polymerase II transcription with ‘hubs’, and proteosome formation proteins with ‘bottlenecks’ were established. In this work, the authors hypothesized that during the evolution of HIV-1, the viral proteins are preferentially attached with central human proteins for direct control and with proteosomal proteins for indirect control over the cellular process. Proteosomal proteins acted as negative regulators for HIV-1 proteins [7, 35]. In [34] it is demonstrated that in early stage of infection, proteosomal proteins regulate some important antiviral host factors like APOBEC3G/F and CD317. In this study, some proteosomal proteins are marked as significantly important based on high bottleneck (or betweenness centrality) scores and linked with important cellular processes.

To investigate the extent to which the proteasomal proteins are associated with crucial and important cellular processes we perform clustering of HIV-interacting proteasomal proteins with Gene Ontology (GO) biological process. For functional annotation clustering, we have used the DAVID bioinformatics resources online service [36]. We find almost 75% proteins among all proteasomal proteins are involved in at least one annotation cluster and are therefore linked to crucial and important biological processes. To show the significance of this association we randomly pick 10 sets of non-proteasomal proteins and perform annotation clustering in similar way. We notice only 4–9.6% of proteins are associated with important biological processes.

In addition, we collect the first and second interacting neighbours of proteasomal proteins and again perform functional annotation clustering with GO biological processes using DAVID. We find almost 78% proteins of the first neighbours and 76% proteins of the second neighbours are associated with important biological and cellular processes. The clustering results of proteasomal proteins as well as their first and second neighbours are given as supplementary files (S1.xlsx, S2.xlsx and S3.xlsx). This study suggests that proteasome proteins are not only responsible for preventing HIV-1 infection at the early stage of infection but also it shows a better connectivity to the important cellular processes because of its high bottleneck scores.

In [34], Dijk et al. find a remarkable over-representation of disease-associated genes among HIV-1 interacting proteins compared with non-interacting ones. In [37], it is reported that degree and betweenness centrality of proteins are poorly correlated with disease association. However, it is tempting to investigate the correlation between the degree and betweenness centrality of disease-associated HIV-1-interacting human proteins. For this purpose we collect list of genes associated with different disorder/disease classes from Goh et al. [37]. They have classified all genetic disorders into 22 primary classes. We study the association of HIV-1 interacting human proteins in these 22 primary disease classes. We plot the degree against the betweenness centrality of all HIV-1 interacting proteins that are associated with specific disease classes. In Figure 7, a total of 22 scatter plots are shown, clearly signifying some sort of relation between degree and betweenness centrality with certain diseases. While for cancer, the correlation appears to be strong (0.9493), for haematological class it is low (0.4840). This indicates that in case of cancer-associated HIV-1 interacting proteins, the ‘hubs’ (proteins with high degree) are also often ‘bottlenecks’ (proteins with high betweenness centrality). Most of the disease classes except ‘Hematological’ and ‘Skeletal’ roughly follow an association between degree and betweenness centrality. In Supplementary Table S1, we have listed all the HIV-1 interacting proteins associated with specific disease classes and their degree and betweenness centrality measured in the corpus of whole human PPI network.

To show whether the correlation patterns between degree and betweenness centrality of HIV-1 interacting disease associated proteins are statistically significant with respect to random set of proteins, we have conducted the Wilcoxon Rank Sum test. For this purpose we take 22 sets of random proteins one for each disease class, from associated HIV-1 non-interacting proteins and measure degree and betweenness centrality in the context of whole human PPI network. We repeat this procedure 300 times to get 300 instances of 22 samples. After that, we compute mean degree and betweenness centrality for each sample. We plot mean degree against mean betweenness centrality for all 22 samples. This is shown in Figure 8. From this figure, it is clear that the correlation is not as strong as it appears for HIV-1 interacting proteins. In fact, the Wilcoxon Rank Sum test shows that there is a significant difference at 5% significance level (P-value = 0.0265).

Interaction types guide regulation of host factor by HIV-1

In section 2 we have seen that the HHPID data set includes 68 unique types of interactions between HIV-1 and host proteins and depending on the regulation type we can subdivide the whole bipartite network into three subnetworks. Similar type of classification of interaction types depending on direction of regulation can be observed in [28] and in [34]. In [28], the interactions are classified into three polar categories (‘positive’, ‘negative’ and ‘neutral’). Positive category involves nine interaction types, viz., ‘activated by’, ‘activates’, ‘enhanced by’, ‘enhances’, ‘stabilizes’, ‘stimulated by’, ‘stimulates’, ‘upregulates by’ and ‘upregulates’, whereas negative category consists of 13 types of interactions, viz., ‘cleavage induced by’, ‘cleaved by’, ‘cleaves’, ‘competes with’, ‘degraded by’, ‘degrades’, ‘disrupts’, ‘downregulated by’, ‘downregulates’, ‘inactivates’, ‘induces cleavage of’, ‘inhibited by’ and ‘inhibits’. The remaining 25 interaction types are denoted as ‘neutral’. To investigate the perturbation of multiple host processes via HIV-1 genes, an analysis is performed to show the involvement of those three classes of interactions in the 10 over-represented biological processes in HIV-1 interactions identified in [9]. It is noticed that for most of the over-represented biological processes’ GO terms, the majority of interactions are mainly ‘positive’ in nature, whereas for immune response, majority of interactions are more ‘negative' in nature. This is obvious because ‘negative’ set contains interactions directed from human to HIV-1 proteins and have an adverse effect on HIV-1 proteins. In [34], almost similar classification of the interaction types is noticed. Here interaction types are classified under two categories: regulatory (e.g. up- and downregulated, regulated by) and activation/inhibition (activates, inhibits, inhibited by), depending on the functional level of interactions. Undirected interactions like ‘binds’, ‘interacts with’ are not considered. Two directed distinct subnetworks are built by considering regulatory interactions in one network and activation/inhibition interactions in the other. A connectivity analysis has been performed for each HIV-1 protein and human protein on both networks. As expected, HIV-1 protein Tat shows high connectivity with host proteins, as it acts as the central transactivator in promoting viral transcription, which affects disease progression by interacting with neighbouring cells after being released to the intercellular medium [38]. Gp120 has almost the same
number of interactions because it is essential in facilitating entry in different cells. On the other hand, the structural proteins P1, P6 and Nucleocapsid as well as unspliced Pol have only a small number of interactions. For performing similar analysis for each human protein they have enriched their HIV-1-human PPI network with interactions from human protein interaction databases BIND [39], BioGRID (http://thebiogrid.org/) and HPRD (http://www.hprd.org/). The interactions between HDFs (HIV dependency factor), which is a class of human proteins that are essential for HIV-1 replication, are considered here. They have constructed two networks: local, which is built by considering the interactions between HIV-1 interacting human proteins or HDFs [13–15, 40], and global, which is constructed by considering the interactions between HDFs and non-HDF proteins. These two networks are then analysed by measuring the degree distributions of HDFs by considering only HDF–HDF interactions and HDFs–HIV protein interactions separately. Besides noting the degree distributions, these two networks are also analysed with respect to the three network centrality metrics: degree, betweenness and eigenvector centrality. They have compared the local HDF network with total human protein interaction network using a one-sided Kolmogorov–Smirnov test [41] and showed that the measured degrees, eigenvector centralities and betweenness scores in the local and global network are not from the same distribution. This suggests that the local network is significantly more central than the global network with respect to the three metrics.

Association of small substructure and strong interaction module with significant host cell subsystems

The biological networks have specifically been found to consist of small recurring patterns, so-called network motifs [42–45]. These building blocks have been extensively used to study the structure and dynamic behaviour of networks. In [34], four types of network motifs were identified in the two subnetworks (regulatory and activation/inhibition) as mentioned in the last section. These significant patterns were identified comparing with 1000 randomized networks, which were created by randomly rewiring the original networks. The significant patterns are shown in Figure 9. The three-node feedback loop motif (shown in Figure 9a) is a pattern where an HIV-1 protein regulates or signals a human protein that in turn regulates/signals another HIV-1 protein. The two-node feedback loop (shown in Figure 9b)
The interactions among proteins signify the nature of feedback pattern in these motifs. For example, two upregulations or two downregulations result in a positive feedback, whereas a negative feedback will be the result of one up- and one downregulation. When two HIV-1 proteins regulate/activate/inhibit one human protein, then co-regulation or co-activation/inhibition patterns are generated (shown in Figure 9c). The three-node ‘clique’ pattern (shown in Figure 9d) consisting of two human proteins and one HIV-1 protein is also identified where the interactions among the proteins have different patterns. A feed-forward type [42–45] (or self-regulatory) motif occurs when two connected HDFs indirectly interact via an HIV-1 protein. An exhaustive analysis of these interaction patterns along with different interaction types may result in understanding of dynamics and structure of significant host cell subsystems that are perturbed during the course of HIV-1 infection.

In Macpherson et al. [46], a nice characterization of interaction patterns is observed that is used to investigate host cellular subsystems and their association in the perturbation of host cellular functions. In this work, they have identified some higher-level biological subsystems and were able to infer the biological importance of these subsystems in terms of HIV-1 replication, host cell perturbation and regulation of the immune response. For identifying key host functions that are entirely involved in HIV-1 infection and replication, they defined a set of host proteins that take part in a common set, or ‘profile’, of

---

**Figure 8.** Scatter plot showing the correlation patterns between degree and betweenness centrality of randomly chosen HIV non-interacting proteins.

**Figure 9.** Four types of network motifs are identified in the two subnetworks (regulatory and activation/inhibition): (a) Three node feedback loop, (b) Two node feedback loop, (c) coregulation/signaling, (d) clique.
HIV-1 interactions. For this they have used BiMax [47] biclustering technique to group the host proteins that encounter similar perturbation during HIV-1 infection. In this work, they retrieved 1434 host proteins and 3939 interactions associated with the specific interaction types from HHPID. A binary interaction matrix containing 1434 rows, 1292 columns and 3939 non-zero values was created corresponding to human proteins, all types of HIV-1 interactions and all HIV-1-human PPIs, respectively. Biclustering of this matrix yielded 1306 biclusters, among which 279 biclusters were identified as significant (P ≤ 0.001) by Monte Carlo simulation. These biclusters were enriched with important sets of human proteins that experienced similar perturbations during HIV-1 infection. The biological significance of host protein sets and their associated interaction profiles defined by the biclusters is assessed according to three measures: PPI network clustering, semantic similarity in terms of shared GO annotation and sequence similarity. They integrated the human proteins from significant biclusters with human PPI network. By this, 38 biclusters were found in which the proteins shared a greater number of interactions, 24 biclusters were found where the proteins form a bigger largest connected component and 38 biclusters were found to be enriched with proteins that have a smaller average shortest path length than that expected by random chance (P ≤ 0.05). These signify that HIV-1 proteins have a strong inclination to interact with host proteins in a similar manner, which share interactions with one another. These also identify the presence of multiple HIV-1 interactions with host protein complexes or other closely associated host network modules. The host proteins in significant biclusters are more similar in terms of their GO annotation than that expected by random chance in all ontologies: molecular function, cellular component (CC) and biological process. These results indicate that HIV-1 interacts in a similar pattern with proteins that have similar GO annotation. It is also observed that the human proteins within the significant biclusters are more similar in their protein sequence than that expected by random chance (P < 0.001). Among all the biclusters, 101 significant biclusters were identified in which human proteins were more similar in their sequences than that expected by random chance. The results show that paralogous groups of host proteins have more inclination to be involved in the same combinations of regulatory and physical HIV-1 interaction. It is observed that several biclusters are enriched with multiple host proteins that show similar interaction profiles among them. By combining these biclusters according to shared information, an overview of HIV-1 interactions with a given set of host proteins is identified.

Some distinguishable patterns of cytokine regulation by HIV-1, such as the largely stimulatory effects of gp120, Tat and Nef; upregulation of TNF-alpha and interleukins 1 and 6; the repressive action of Vpr and gp160; and downregulation of interleukin 2 and interferon-c are identified. Two HIV-1-host PPI networks are also created in which the human host is depicted as a series of cellular subsystems and the interactions of HIV-1 proteins with these subsystems are investigated to find some recognizable patterns of interaction in these networks.

Using a neighbourhood joining method, identified biclusters are combined. A tree is constructed using a distance measure based on the overlaps between host proteins constituting the biclusters. The tree is partitioned into different sections according to the over-representation of the GO terms of the host proteins in the combined biclusters. Thirty-seven biological subsystems are identified to be associated with the created partitions. Associating the biclusters in this way with 37 biological subsystems, strong associations between biclusters and specific subsystems are identified.

It will be interesting to see the association between biclusters that consist of human proteins having the same regulation type with HIV-1 and the aforesaid 37 biological subsystems. For this purpose we use biclustering technique to define biclusters in each of the three subnetworks (‘positive’, ‘bidirected’ and ‘negative’) identified in HIV-human PPI network by considering direction of regulation of each interaction. A total of 110 biclusters (69 from subnetwork 1, 21 from subnetwork 2 and 20 from subnetwork 3) are identified using Bimax [48] algorithm. Figure 10 shows overlap of subsystems involved in the biclusters of three subnetworks, viz., positive, bidirected and negative. It appears from the figure that most of the subsystems do not show propensity to be involved in some specific interaction types. Only six subsystems, viz., ‘Glutamate receptor activity’, ‘protein phosphatase type 2A complex’, ‘DNA integration’, ‘lactate dehydrogenase activity’, ‘adenylate cyclase activity’, ‘cytoskeleton’, four subsystems, viz., ‘ubiquitin’, ‘heat shock 70kDa protein’, ‘glucosidase activity’, ‘miRNA transport’ and one subsystem, viz., ‘DNA helicase activity’ show some sort of inclination to be involved with specific regulation types like ‘positive’, ‘negative’ and ‘bidirected’, respectively.

To present a distilled representation of the relationship between over-represented biological subsystems mediated through regulation types, we have constructed three bipartite networks specific to each of the three regulation types. Figure 11a has total 69 nodes, each of which corresponds to one bicluster and 27 nodes representing the biological subsystems associated with those biclusters. The small yellow nodes represent biclusters, whereas red nodes stand for the subsystems. To know to what extent each subsystem is associated with each bicluster, we vary the node size of the subsystems proportionally to the associated proteins with all the biclusters. The edges in this network represent bicluster–subsystem association that alignates the subsystems in a regulation-specific manner. The edge width is proportional to the proportion of proteins in the biclusters associated with the subsystems. It can be observed that most of the subsystems solely associated with negative regulation have a clear role in immune response in the host cell. For example, it is experimentally verified that synthesis of glucosidase show antiviral and immunosuppressive activity of HIV-1 by inhibiting HIV-1-induced syncytia formation and lymphocyte proliferation [48, 49]. In [50], it is also verified that NEDD8 (a small ubiquitin-like protein) affects the restoration of Vif-mediated degradation of the host restriction factor APOBEC3G (A3G), and this presents a novel antiretroviral therapeutic approach and enhances the ability of the immune system to combat HIV.

Recently, Maulik et al. [51] proposed a multi-objective biclustering technique for extracting strong interaction modules in a weighted HIV-1-human PPI network. A strong interaction module consists of a set of viral and host proteins, which are highly interacting with each other. The human proteins involved in those strong interaction modules are likely to be considered as important gateways for HIV-1 proteins and may be regarded as potential drug targets. In this work, the HIV-1-human protein interaction network is realized as a bipartite graph where two sets of nodes represent HIV-1 proteins and human proteins, respectively, and edges denote the interactions. The network is weighted by interaction strength predicted by Tastan et al. [52]. Each ‘quasi-biclique’ (dense bipartite subgraph) represents a strong interaction module in this weighted bipartite network. The main issue in this work is that it relaxes the stricter
condition for forming bicliques [53] by introducing ‘quasi-biclique’ that are considered as almost bicliques in HIV-1-human bipartite protein interaction network. Searching is carried out over three different objectives for finding the quasi-bicliques in HIV-1-human bipartite protein interaction network. The resulting strong interaction modules (or ‘quasi-bicliques’) are then analysed to investigate the involvement of dense sub-networks during HIV-1 infection cycle. Some important human proteins that are responsible for early and late stages of HIV-1 infection were found in the extracted ‘quasi bicliques’. For example, at least one occurrence of kinase protein is observed in all the identified quasi bicliques. Beside the Kinase family proteins the bicliques are also enriched with CD4, chemokine receptor CXCR4, AP2B1 like proteins. Gene Ontological analysis over the human proteins identified in the ‘quasi-bicliques’ reveal several biologically homogeneous strong interaction modules.

Incorporating other data sources with HHPID in the analysis of HIV perturbation in host cell line

Besides the study of the network built from HHPID data set, the key to deriving new understanding from the HHPID lies in its integration with other data sources. For example, Chen et al. [54] analysed the association between HIV-1 infection and human pathways of diseases that may lead to the identification of common drug targets for viral infections and other diseases. In this work, they identified links between HIV-1 infection and other human pathways of disease through several approaches. They primarily investigated the overlap of human genes involved in AIDS and other disease pathways. They gathered 220 human pathways in disease from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/) and statistically compared with three sets of HIV-1 host factors identified from three systematic screening studies [13–15]. The evaluation of associations between HIV-1 host factors and KEGG pathways is based on (i) overlap of human genes involved in AIDS and other pathways, (ii) co-expression profiles and (iii) common interaction partners in a human PPI network. In particular, the human pathways are ranked based on these evaluation scores. The top 10 KEGG disease/pathways that emerge from the final ranking list are ‘Pancreatic cancer’, T-cell receptor signalling pathway’, ‘Acute myeloid leukaemia’, ‘B cell receptor signalling pathway’, ‘small cell lung cancer’, ‘chronic myeloid leukaemia’, ‘Adipocytokine signalling pathway’, ‘toll-like receptor signalling pathway’, ‘chemokine signalling pathway’ and ‘apoptosis’. Figure 12 shows the proportion of HIV-1 interacting human proteins that are associated with these pathways. It can be seen that most of the HIV-1 interacting proteins are involved in each pathway. So, it is important to investigate whether these HIV-1 interacting proteins have different characteristics from HIV-1 non-interacting proteins that also appear in these pathways. For this, in Figure 13 we plot a scatter diagram showing the degree versus betweenness centrality of HIV-1 interacting proteins and HIV-1 non-interacting proteins involved in each of the pathways. Interestingly, we find a significant correlation between degree and betweenness centrality of HIV-1 interacting proteins than HIV-1 non-interacting counterparts. It is also noticeable that in most of the pathways, HIV-1 interacting proteins have more tendency of being ‘hubs’ or ‘bottlenecks’ than the non-interacting ones. So, it can be stated that HIV-1 makes some sort of perturbation in the disease-related pathways that affect the topological properties of the pathway-related proteins in the context of the whole human PPI network.
Figure 11. Association between 37 subsystems and biclusters in three different subnetworks, viz., (a) positive, (b) bidirected and (c) negative. Edge width represents proportion of proteins in biclusters that are matched with subsystems. Node size is proportional to the number of matched proteins with all biclusters. Blue nodes indicate that these subsystems are solely associated with corresponding regulation specific biclusters. A colour version of this figure is available at BiB online: http://bib.oxfordjournals.org.

Figure 12. Proportion of HIV interacting proteins involved in 10 top-ranked pathways reported in Chen et al. [54].
In another work, Bandyopadhyay et al. [55] analysed a time-course gene expression data for identifying active pathways of connected proteins that were significantly under/over-expressed in the different stages of latent HIV-1 replication cycle. The differentially expressed genes were traced in the human protein interaction data set [HPRD] [56] for creating an integrated network. Some significant ‘activated modules’ of connected proteins at different stages of virus reactivation cycle were identified through expression clustering. In summary, this work permitted the identification of several human proteins whose expression levels were associated with either the early induction or the latent phase of viral replication.

In another study [57], several data sources are integrated with the existing information catalogued in the HHPID to find out the activity of miRNAs through the HIV-1 regulatory pathway in human. For this they used two TF-miRNA regulation information from two data sources [58, 59] and one putative database [60]. An exhaustive search was conducted to find maximal bicliques from the HIV-1-human bipartite network constructed from the HHPID database. These bicliques represent strong interaction modules consisting of a set of HIV-1 proteins and a set of human proteins. These modules are further investigated to uncover their involvement in possible oncogenic pathways. The significant modules discovered in this study are established as gateways to transmit the immunodeficiency signal within the complex regulatory network of non-coding and coding genes within an organism.

Figure 13. Scatter diagram showing the degree versus betweenness centrality of HIV interacting protein and HIV non-interacting proteins involved in 10 top-ranked pathways reported in [54]. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

A review on HIV-1-human PPI prediction techniques

PPIs are regarded as the main biochemical reactions in cells, which determine different biological processes, organized in the cell. There exit several literatures that assess different methodologies used for predicting PPIs [61–63]. Most of the approaches predominantly focus on the prediction of PPIs in a single organism such as yeast or human. Here we discuss several computational methods that concentrate on the HIV-1-human PPI prediction. We also evaluate the applicability of different methods and perform different analysis to compare those methods at the system level.

Computational methods for prediction of possible viral–host interactions are one of the major tasks in PPI research for antiviral drug discovery and treatment optimization. Predicting PPIs between viral and host proteins has contributed substantial knowledge to the drug design area. Recently, PPI prediction has been regarded as a promising alternative to the traditional approach to drug design [64]. Having knowledge about the interaction pattern provides a great opportunity to understand pathogenesis mechanisms, and thus supports the development of drugs.

There are different approaches available in the literature that exploit different methodologies for predicting PPIs between HIV-1 and human proteins. Here, we have classified these approaches into three categories depending on the methodologies, which have been used in those literatures. In the first category, we collect the works that use classification-based approaches for detection of PPIs. In the second category, we describe a work that mainly depends on the structural similarities of HIV-1 and human proteins. Here we also describe a work that uses the short eukaryotic linear motifs (ELMs) [65] on HIV-1 proteins and human protein counter domains (CDs) information for predicting interactions between HIV-1 and human proteins. Recently association rule mining techniques are gaining popularity in this domain and are established to be a useful alternative in the field of PPIs prediction. We mark these works as the
third category. The classification of these approaches is shown in Table 2.

Category-1: classification-based approaches

In [52], a supervised learning framework is presented for predicting PPIs between HIV-1 and human proteins. It was the first attempt to predict the global set of interactions between HIV-1 and human host cellular proteins. In this work, the task of predicting PPIs is formulated as a binary classification problem where each protein pair belongs to either the interaction or non-interaction class. A Random Forest Classifier (RF) is trained with a rich feature set derived from several biological information sources. Total 35 features are recognized based on human intra-PPIs and other information sources. Some features precisely capture the information from HIV-1 human protein pairs, while others are related to only human or HIV-1 proteins, and some are derived from the human interactome. As these extracted features are noisy and redundant, therefore a RF [73] has been used for the classification task. RF classifier has already been established to be useful in predicting intra-species PPIs [74–76]. It is also suitable for handling the scenarios where feature set is noisy and redundant. For building the trees in RF in [73], a node is chosen to be split, when the attribute causes highest decrease in the GINI index. All the features are ranked depending on the Gini importance of the RF classifier. The topological properties like degree and betweenness centrality of the graph, and the GO neighbour similarity features got the top rank. Degree is coming out here to be a top-ranked feature. It is quite expected because degree is the property of a single node and provides a good prior interaction probability for the prediction purpose. The other top-ranked features with respect to the Gini importance are pairwise GO similarity, clustering coefficient, GO neighbour function, location features etc. It is pertinent that Gene Ontological features come out to be top-ranked features. It is possibly because of the importance of GO similarity measures or other GO-related features that contribute greater knowledge towards interaction prediction [77, 78]. The authors of [52] have extended their work in [66] by integrating a semi-supervised multi-tasking approach to improve the accuracy of the previous predictive model. In this work, they reduced their feature set but while improving the accuracy by considering ‘weakly labelled’ PPIs. The ‘weakly labelled PPIs’ are not experimentally confirmed PPIs, but are likely to have interaction relationships. Here the motivation of taking these weakly labelled PPIs is the insufficiency of confirmed labelled PPIs. The PPIs in the HHPID [10], which are retrieved from different scientific literature, may not have enough evidence of being true positive labelled interaction. But these interactions may be a possible candidate to contribute greater knowledge towards the PPI prediction. To incorporate the ‘weakly labelled PPIs’ information with ‘true positive labelled PPIs’ the authors in [66] have used the multi-task learning framework. For selecting the positive labelled PPIs and partially labelled PPIs (or weakly labelled PPIs), they collected experts’ opinions for improving the quality of the data set. The recruited HIV experts contributed their knowledge to annotate the interactions of HHPID in truly positive label and partially positive label. The 158 pairs, annotated by HIV-1 experts, are considered here as ‘gold standard positive pairs’ and the remaining 2119 pairs are labelled as partially positive. For choosing the negative set, approximately 16 000 random protein pairs are collected excluding the pairs existing in the HHPID. To compensate the effect of random sampling of negative training set, 5-fold cross validation (CV) with 20 repeated CV runs are performed to obtain the average performance scores.

In another study, Dyer et al. [67] proposed a supervised classification technique using Support Vector Machine (SVM) with a linear kernel for predicting PPIs between HIV-1 and human proteins. For choosing negative examples, they also used random pairing of HIV-1 and human proteins. In doing so, they ensured that no randomly generated protein pair was already known to interact. They generated three sets of negative examples containing 25, 50 and 100 times the number of the positive examples and measured the performance of SVM. The three types of features used in this work are domains, protein sequence k-mers and properties in the intra-species human PPI network.

In [79], a probability weighted ensemble transfer learning model is proposed to predict PPIs between HIV-1 and human proteins. Here, the homolog GO information is incorporated to cope up with the three major problems of PPI prediction, viz., data scarcity, data unavailability and negative data sampling [79]. SVM is adopted here as the individual classifiers of the ensemble model. In this work it is demonstrated that homolog GO information is effective to enrich or substitute the target GO information.

Irrespective of all the classification techniques reviewed here, the task of choosing the negative examples is generally observed as a common problem. In most of the cases for choosing non-interacting pairs, a set of protein pairs are extracted at random from the set of all protein pairs. However, there is no evidence that these chosen pairs do not interact at all. As a result, the classification task remains sensitive towards the chosen negative samples. In [40], the negative to positive ratio is kept as 100:1, and this ratio is optimized in each CV step in the training phase. In [67], the negative to positive ratio was varied as 25:1, 50:1 and 100:1, and using each of these ratios the performance of SVM classifier is measured under different feature sets. In [66], a remedy of this problem was observed by performing 5-fold CV with 20 repeated CV runs for compensating the effect of random sampling of negative training set. Here a clear separation is maintained for the negative labelled data and partially positive labelled data set. A classifier is trained to distinguish partial positive examples from negative examples. This refinement of data set leads to a significant improvement in identification of interacting protein pairs.

Category-2: structural similarity and motif sequence-based approach

In Doolittle et al. [68], the significant amount of protein structural information available in the Protein Data Bank are used for predicting PPIs between HIV-1 and human proteins. The use of protein structure information integrated with the documented PPIs is established to be an alternative way for the prediction of possible protein interactions [80–82]. The motivation of using the structural information is that given a set of proteins with known structure and interaction partner, the proteins that show similar structure with these given proteins are likely to interact with those interaction partners. In [68], this argument is validated in context of HIV-1 and human proteins. Based on the protein structural similarity, an interaction map between human and HIV-1 is generated. For this, interactions between ‘HIV-similar’ (i.e. structurally similar to some HIV protein) human proteins and ‘targets’ (i.e. the human proteins interacted with ‘HIV-similar’) are identified and an association among the interactions between HIV-1 protein and the ‘target’ proteins is established. The predicted interactions are filtered
proteins that act as a competitor of HIV-1 protein to interact with a H1 protein via the conserved ELM. Therefore, if H1 has a CD, then it may use this CD to interact with an ELM present on both H2 and HIV-1 proteins.

**Category-3: association rule mining-based approaches**

In Mukhopadhyay et al. [70], an attempt was made to predict interactions based on association rule mining technique. Here the well-known Apriori [86] algorithm is used for generating association rules between human proteins interacting with HIV-1 proteins. Some novel interactions are then predicted using those rules. Besides the prediction task, the discovered rules describe an association between different human proteins in the context of HIV-1 interaction. The main advantage of using the association rule mining [72] technique for prediction purpose, is that it does not need any negative samples for prediction unlike classification-based approaches, where predictions are highly

---

**Table 2. Summary of the works involved in the PPI prediction field**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Operating principle</th>
<th>Approach</th>
<th>Number of predicted interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tastan et al. (2009) [52]</td>
<td>Mainly supervised classification-based approach where a RF is trained and tested against an extensive rich feature set combining co-occurrence of functional motifs and their interaction domains and protein classes, GO annotations, post-translational modifications, tissue distributions and gene expression profiles, topological properties of the human protein in the interaction network and the similarity of HIV-1 proteins to human proteins known binding partners.</td>
<td>Classification-based approach</td>
<td>3493</td>
</tr>
<tr>
<td>Qi et al. (2010) [66]</td>
<td>A semi supervised multi-tasking framework is introduced for predicting PPIs from a wide range of labelled (experimentally validated) and partially labelled (there are no experimental evidence of those interactions) data set.</td>
<td>Classification-based approach</td>
<td>2084</td>
</tr>
<tr>
<td>Dyer et al. (2011) [67]</td>
<td>A SVM classifier with linear kernel is used for exploring the predicted interactions in HIV-1-human system. Here domain profiles, protein sequence k-mers and properties of human proteins in PPI network are incorporated in feature set.</td>
<td>Classification-based approach</td>
<td>506 (with PE:NE ratio 1:50)</td>
</tr>
<tr>
<td>Doolittle et al. (2010) [68]</td>
<td>Protein structure information is integrated for prediction of PPIs. The main idea behind this technique is that given a set of proteins with defined structures and associated interactions, proteins with similar structures or substructures with them will tend to share the same interaction partners.</td>
<td>Structural similarity-based approach</td>
<td>884 (using literature filter)</td>
</tr>
<tr>
<td>Evans et al. (2009) [69]</td>
<td>Interactions are mediated by short ELMs on HIV-1 proteins and human protein CDs known to interact with these ELMs. The decision on which human protein may be targeted by HIV-1 is guided by taking pair of human proteins that may interact via a motif conserved in HIV-1 and the corresponding interacting protein domain.</td>
<td>Motif sequence-based approach</td>
<td>4542</td>
</tr>
<tr>
<td>Mukhopadhyay et al. (2010) [70]</td>
<td>Association rule-based prediction of novel PPIs. The well-known Apriori algorithm is used for finding association rules among the HIV-1 and human proteins. Some novel PPIs are also predicted using those rules.</td>
<td>Association rule mining-based approach</td>
<td>22</td>
</tr>
<tr>
<td>Mukhopadhyay et al. (2012) [71]</td>
<td>An association rule mining technique based on biclustering is used for predicting PPIs. Predictions are defined from the association rules that are generated using a novel biclustering-based association rule mining technique.</td>
<td>Association rule mining-based approach</td>
<td>180</td>
</tr>
<tr>
<td>Mondal et al. (2012) [72]</td>
<td>An integrated approach called FIST is proposed here for procuring association rules among HIV-1 and human proteins. Here FIST is used to coalesce biclustering and association rule mining technique in one process.</td>
<td>Association rule mining-based approach</td>
<td>28</td>
</tr>
</tbody>
</table>
dependent on the availability of the biologically validated negative samples. In [71], the authors have extended their work by using a biclustering technique for discovering association rules between both human and viral proteins. Two binary matrices of human and viral proteins, HV of size 1432 × 19 and its transpose matrix VH of size 19 × 1432 are built where an entry of ‘1’ denotes the presence of interaction between the corresponding pair of human and HIV-1 proteins, and an entry of ‘0’ represents the absence of any information regarding the interaction. For each of the matrices, a row is considered as a transaction and a column represents an item. In each transaction, the items for which the corresponding value in the matrix is 1 are considered to be purchased in that transaction. Now, the columns of all-1 biclusters, which are extracted from those matrices using Bimax algorithm, represent frequent closed itemset. The frequent itemsets are then used to generate association rules among human proteins and viral proteins separately. These rules basically guide the prediction task.

In another approach [87] proposed by Mukhopadhyay et al., regulatory interactions are predicted by using association rule mining technique. Here all the interactions are partitioned based on the interaction type and direction of regulation. After partitioning, three different regulatory networks are formed based on the HIV–1–human interaction information. The interactions are predicted by considering each of the networks separately. As here each interaction has a specific interaction type and direction so, all the predicted interactions also inherit these information.

The approach proposed in [88] also used the association rule mining technique for extracting frequent closed itemset for predicting PPIs between HIV–1 and human proteins. They proposed an integrated approach called FIST (Frequent Itemset Mining using Suffix Tree) that combines biclustering and association rule mining technique in a single process. The main feature of FIST is that it efficiently mines frequent closed itemsets, generates minimal non-redundant cover of association rules and generates hierarchical conceptual biclusters in one process. It was applied on the HHPID database for predicting some novel interactions.

All these three approaches are apparently correlated with respect to the methodology they have used, but different techniques are used to mine the association rules. For example, the first work of Mukhopadhyay et al., the Apriori algorithm, is used for the purpose of finding association rules. But in case of low minimum support, large number of frequent itemsets is generated, which results in large computational complexity. Generation of such a large number of frequent itemsets takes huge time. In this context the concept of frequent closed itemset [89, 90] is used in [71] as an extension of first work. An itemset is closed itemset if none of its proper supersets have the same support value. Here BiMax [47] biclustering algorithm is used to generate all maximal biclusters. As the columns of maximal biclusters represent a closed itemset, so all extracted biclusters satisfying a predefined minimum support condition provide the set of frequent closed itemsets. However, in Mondal et al. [88], a frequent closed itemset framework is proposed for efficiently mining closed itemset. It is also used to find out the relationship between annotations and interactions by extracting association between proteins and features.

Analysis of the predicted interaction sets

In this section, we have analysed all predicted interaction sets identified in different literature. In the following subsections we make comparisons of those predicted interaction sets in different perspectives.

Detecting overlaps between predicted interactions

In Figure 14 we analyse the number of interactions predicted by Tastan et al. [66], Doolittle et al. [68], Dyer et al. [67] and Mukhopadhyay et al. [71]. Note that these studies use different methodologies and different data sets for learning how to predict the interactions. However, the predictions, as analysed in Figure 14, have been obtained on the same set of 19 HIV–1 proteins and 1432 human proteins, where each method makes predictions on all possible (19 × 1432) interactions. From this figure, it can be seen that there is no significant overlap among the interactions predicted by these four studies. Similar outcome has also been noticed in Figure 5 where the overlaps among the predicted HDF sets are low. Doolittle et al. exploited structural similarity information of HIV–1 and human proteins for prediction purpose. Therefore, the human protein that does not show structural similarity with any HIV–1 proteins cannot be included in the possible prediction set. Moreover, they used HPRD and PIG database for collecting information about interactions, Dali and PDB database for acquiring information about structural similarity and HHPID data set for validating the predictions. Although Tastan et al. and Mukhopadhyay et al. used HHPID data set for prediction purpose, the main drawback of Mukhopadhyay et al. technique is that it cannot predict any interaction whose protein pairs are not included in a maximal biclue. The method proposed by Tastan et al. produces all possible pairs of interactions and is able to compute prediction score of each possible interaction pair. However, they did not validate their predictions by exploring literature survey. In Figure 13, slightly larger amount of overlap is noticed among Dyer et al. and Tastan et al., possibly because Tastan et al. incorporated PPI network properties in their feature set, which also serve as a key predictor in Dyer et al.

Over-representation of 19 HIV proteins in the four predicted interaction sets

We perform a study to show which HIV–1 proteins are over-represented in the predicted interaction set in each of the four studies. In Figure 15 we see that in most of the predicted interaction sets, HIV–1 protein TAT is significantly over-represented, possibly because of its essentiality for efficient transcription of the viral genome [91, 92]. However, a remarkable under-representation of TAT is found in the predicted interaction set.
discovered by Doolittle et al. This is possibly because of the lack of structural similarity between Tat and human proteins and small number of existing 'Tat-similar' human proteins. The predicted set of Doolittle et al. is enriched with transmembrane glycoprotein gp.41 but surprisingly surface glycoprotein gp.120 is not recognized in this method. So their method is slightly biased towards the availability of the structural similarity between HIV-1 proteins and human proteins. But when we consider the predictions extracted by association rule-based approach developed by Mukhopadhyay et al. we see different things. Here the interactions involving surface glycoprotein gp.120 are significantly recognized but gp.41 remains unrecognized. Among the four methods, a small proportion of F7/nucleocapsid protein and retropepsin are recognized in the rule-based approach of Mukhopadhyay et al. The predicted interaction set of Tastan et al. is more or less enriched with most of the HIV-1 proteins. A rich feature set including different biological and network properties used for prediction may be a possible reason for that. In the predicted set of Dyer et al. we see there are smaller numbers of HIV-1 proteins that are involved in the interactions with human proteins with respect to that of Tastan et al.

HDF enrichment in the four predicted interaction sets

We have also studied these four works in the context of enrichment of HDFs in their predicted sets. For this we gathered 275 HDFs from the study done by Brass et al. [13], 296 HDFs from the study done by Konig et al. [14] and 375 HDFs from the study done by Zhou et al. [15]. There were 908 unique HDFs in the union of these sets. We measure the proportion of overlaps between the predicted set of Tastan et al., Doolittle et al., Dyer et al. and Mukhopadhyay et al. with these HDFs set. In Figure 16 the overlaps between the predicted interaction sets for each of these four works and the three HDF sets are shown. For this, we intersect the human proteins in the predicted interaction set with three HDF sets individually and measure the proportion of the human proteins that belong to these HDF sets. We find that a large proportion of human proteins in the predicted set of Doolittle et al. belongs to each of the experimentally predicted HDF sets. Figure 17 also shows the proportion of overlapped human proteins in the interaction sets that belong to the union of all three HDF sets.

Overall assessment of all predicted sets using conformal approach

As the predicted interaction sets are dependent on the corresponding methodologies used for prediction, it is somewhat unfair to draw conclusion regarding the superiority of any method based on simply comparing them. Still to get an overview of the quality of the predicted sets, we have made an attempt to compare the predicted sets. For the purpose of a comprehensive comparison, we take the conformal prediction approach as proposed in [93]. It uses 35 features collected from Tastan et al. [52] and assigns a confidence level to each predicted pair. We use this rich feature set because it integrates most of the physical properties of the proteins. Moreover, for all possible interaction pairs the values of corresponding features are available in [93]. Usually conformal prediction is more concerned with two or more classes. However, in [93] conformal prediction is used to deal with one defined class, which consists of experimentally validated interaction pairs. Here the Non-Conformity Measure deals with the feature sets and a P-value is assigned to individual predicted interaction to validate how likely it is with respect to the previously defined interactions. More is the P-value of an interaction, more it is likely to occur. Here we find P-values of all predicted interactions of each of the four studies and plot the proportion of interactions against the P-values. Figure 18 shows column plots of predicted interactions with P-values for each of the four studies, separately. From this figure it is clear that the prediction sets of Tastan et al. and Doolittle et al. have interactions that are more likely than the interactions predicted by Mukhopadhyay et al. and Dyer et al. (over 60% of interactions of Tastan et al. and Doolittle et al. have P-value ≥ 0.6, but for Dyer et al. and Mukhopadhyay et al. it is 33.05 and 56.50%, respectively). Note that this is expected because the 35 features used in the conformal prediction were collected from Tastan et al. [52]. For lack of better alternative, we use this approach for comparing the predictions.

Conclusion

We provide an overview of computational methods on the analysis of HIV-1-human PPI network and the prediction of novel PPIs between HIV-1 and human proteins. We have extensively surveyed a number of works, from both analysis and prediction
aspects. The mostly used database, which has been extensively used for these studies, is the HHPID database. Some of the studies refined this curated data set before using it in their methods. For example, Qi et al. [66] provided an expert’s view of the interaction data set for selecting the positive and negative samples before applying classification task. Similar type of rectification of this data set is noticed in several other studies. However, the availability of this data set of HIV-interacting human

Figure 16. Proportion of overlaps between human proteins in the predicted set of four studies and the three experimentally verified HDF sets. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

Figure 17. Proportion of overlaps between human proteins in the predicted set of four studies and union of all HDFs gathered from Brass et al. [13], Konig et al. [14] and zhou et al. [15].

Figure 18. Bar plots show proportion of predicted interactions having specific P-values. X-axis represents ranges of P-values and Y-axis represents proportion of interactions predicted by the four studies viz., Doolittle et al., Mukhopadhyay et al., Dyer et al. and Tastan et al.
proteins raises the possibility of identifying putative novel drug targets by in silico methods.

The approaches designed for network analysis used different computational and statistical methods and some topological properties of network to uncover the inherent relationship between HIV-1 and human proteins. Among all the works we have surveyed here, biclustering technique is extensively used in network analysis as well as for PPI prediction. In the network analysis field, we find a gradual change in the analysis technique. Former works are fully devoted to the analysis of topological properties of the HIV-1-human PPI network. Recently analysis of bicliques or biclusters extracted from the HIV-1-human PPI network sheds light on the modular structures of the network. Several techniques we have reviewed here pertaining to network analysis, have taken this approach.

The techniques for interaction prediction are also regarded as important resources for discovering interesting relationships between HIV-1 and human proteins. Although the works in this category primarily predict some novel interactions between HIV-1 and human proteins, the methods used for this purpose make use of divergent techniques. Classification-based methods are widely used for prediction. For example, Dyer et al. [67] made use of SVM-based classifier, whereas RF is used in Tastan et al. [52] and Qi et al. [66]. However, all the classification techniques suffer from insufficiency of negative examples. To compensate for this problem, different techniques are used. For example, in Dyer et al., negative to positive ratio is varied and results are collected by feeding them in SVM classifier with different feature sets. Another example can be found in Qi et al. where 16 HIV experts are recruited for collecting experts view on interacting and non-interacting pairs of HIV-1 and human proteins. Among all the prediction methods studied in this assessment, only Doolittle et al. used the protein structural information for predicting PPIs between HIV-1 and human proteins. We found significant overlap between the predicted set of Doolittle et al. and HDF set identified in large-scale siRNA screening. Some of the other approaches like Mukhopadhyay et al. and Mandal et al. do not show a remarkable overlap with the HDF sets. These approaches made use of association rule mining technique for prediction purpose. But the rule mining has certain limitations for predicting interactions. It cannot predict the interactions among proteins that are not a part of some biclique. On the other hand, Dyer et al. and Tastan et al. were able to give prediction scores to all possible interaction pairs. However, as their predicted interactions have little overlap with HDF sets, it is hard to believe that these interactions surely exist in reality. However, the conformal approach we have studied here also gives confidence to a new interaction based on the feature set derived by Tastan et al. As this feature set is rich enough, the interactions with a high P-value are reliable with some confidence. It also may be noted that the predictions in Doolittle et al. are associated with low P-values but high overlap with HDFs. This contrasts with the results of Tastan et al., indicating that the two methods are probably predicting two different subsets of true PPIs. One of the aims of reviewing all these methodologies and examining their predicted interaction sets is to highlight the common ones, which are supported by most of the techniques. Though this number is small, but there is enough reason to believe these interactions to be real as most of the methodologies agree on them.

In this article, we completely focus on the computational effort that has been made for analysing HIV-1-human PPI network and predicting interactions between HIV-1 proteins and human proteins. A brief discussion on some experimental approaches for HIV-1-human PPI identification is provided in the supplementary file (experimental-technique.docx). Similar type of review may be done on other type of viruses and pathogens. If sufficient information is available on a viral–host interaction network, these can be analysed accordingly.

Although the works we reviewed here suffer from several problems and deficiencies, there are reasons to believe that these in silico approaches will soon be of immense importance for understanding infectious diseases in general and in particular for the development of drugs and vaccines to counter viral replication or the pathogenic effects of infection. There are more to discover from the systems biology perspective that focus on data integration from different sources and complex interplay between virus and host proteins in the virus-host PPI network. A detailed understanding of this intricate network and a precise study of its properties will yield novel insights into the development of new therapeutics and potentially new intervention strategies.

**Supplementary data**

Supplementary data are available online at http://bib.oxfordjournals.org/

**Key Points**

- The computational approaches for analysing HIV-1-human PPI network and predicting PPIs in it can broadly be classified in two categories, viz., network analysis and PPI prediction.
- In network analysis field HIV-1-human PPI network is analysed using different computational and statistical methods in which a gradual change is noticed from the topological analysis to the modular analysis of the network.
- The approaches in the PPI prediction field can be classified into three categories, viz., classification-based, protein structural similarity-based and association rule mining-based approaches.
- Due to the uncorrelated methodologies used in the prediction field, poor overlap is observed among the predicted sets.
- More study is required for integrating information from different data sources for development of new therapeutics.

**References**


84. Shelton H, Harris M. Hepatitis c virus ns 5 a protein binds the sh 3 domain of the fyn tyrosine kinase with high affinity: mutagenic analysis of residues within the sh 3 domain that contribute to the interaction. Virology 2008;5:24.


