Journal of Bioinformatics and Computational Biology
Vol. 11, No. 2 (2013) 1230001 (26 pages)

DOI: 10.1142/S0219720012300018

Imperial College Press
www.icpress.co.uk

PROGRESS IN COMPUTATIONAL STUDIES OF HOST–PATHOGEN INTERACTIONS

HUFENG ZHOU[*,‡], JINGJING JIN[†] and LIMSOON WONG[†,§]

[*]*NUS Graduate School for Integrative Sciences & Engineering*
National University of Singapore, Singapore 117456

[†]*School of Computing, National University of Singapore*
Singapore 117417
[‡]*zhouhufeng@nus.edu.sg*
[§]*wongls@comp.nus.edu.sg*

Received 18 June 2012
Revised 16 August 2012
Accepted 23 August 2012
Published 22 October 2012

Host−pathogen interactions are important for understanding infection mechanism and developing better treatment and prevention of infectious diseases. Many computational studies on host−pathogen interactions have been published. Here, we review recent progress and results in this field and provide a systematic summary, comparison and discussion of computational studies on host−pathogen interactions, including prediction and analysis of host−pathogen protein−protein interactions; basic principles revealed from host−pathogen interactions; and database and software tools for host−pathogen interaction data collection, integration and analysis.

Keywords: Host−pathogen interaction; protein−protein interaction; PPI; infectious diseases.

1. Introduction

Infectious diseases are among the leading causes of death worldwide. Host−pathogen interactions are crucial for better understanding of the mechanisms that underline infectious diseases and for developing more effective treatment and prevention measures. While host−pathogen interactions take many forms, in this review, we concentrate on protein−protein interactions (PPIs) between a pathogen and its host. This review consists of the following parts: (i) host−pathogen PPIs prediction; (ii) basic principles derived from the analysis of known host−pathogen PPIs; (iii) host−pathogen PPIs analysis and assessment; and (iv) host−pathogen interaction data collection and integration.

Several approaches have been proposed to computationally predict host−pathogen PPIs. There has also been progress on analyzing and assessing the quality of the inferred host−pathogen PPIs. This has led to cataloging of PPI data that can be further analyzed to understand the impact of these interactions (especially on the

host) and to decipher underlying disease mechanisms. Approaches developed for predicting host−pathogen PPIs can be broadly categorized into homology-based,[1−5] structure-based,[6−8] domain−motif interaction-based approaches,[9,10] as well as machine-learning−based approaches.[11−13] These approaches can also be combined and used together in some studies to improve prediction performance. These approaches are reviewed in Sec. 2.

An analysis of experimentally verified as well as manually curated host−pathogen PPIs have led to a number of observations. These observations include the topological properties of targeted host proteins and structural properties of host−pathogen PPI interfaces. These observations are discussed in Sec. 3.

Approaches for assessing and analyzing host−pathogen PPIs can be categorized into assessment based on gold standard PPIs,[6,7,10−13] functional information analysis [Gene Ontology (GO),[5−8,10,11] pathways,[5,10,14,15] gene expression data,[2,5,6] RNA interference data[7,8,10−13]], localization information analysis (protein subcellular localization,[1−5] co-localization of host and pathogen proteins[7,8]), related experimental data analyses[8,11,13] and biological case studies and explanations.[2−4,6−8,12] Some of these assessment approaches can also be used as filtering strategies for pruning host−pathogen PPI prediction results. These approaches and the outcome of the analysis are reviewed in Sec. 4.

Host−pathogen PPIs curated from primary literature are usually facilitated by text-mining techniques.[16,17] With more host−pathogen PPI data available from literature curation and experiments, there are strong needs for data collection and integration facilities that can provide comprehensive storage, convenient access and effective analyses of the integrated host−pathogen interaction data. The development of software and database tools dedicated to host−pathogen interaction data collection, integration and analysis are also very prominent. Integration of host−pathogen interaction data is not confined to PPI data. Other related data — like pathogen virulence factors, human-diseases−related genes, sequence and homology information, pathway information, functional annotations, diseases information, literature sources, etc. — are also being integrated into several databases. These databases[16−25] and softwares[26] are reviewed in Sec. 5.

2. Host−Pathogen PPIs Prediction

Host−pathogen PPIs play an important role between the host and pathogen, which may be crucial in the outcome of an infection and the establishment of disease. Unfortunately, experimentally verified interactions between host and pathogen proteins are currently rather limited for most host−pathogen systems. This has motivated a number of pioneering works on computational prediction of host−pathogen PPIs. These works can be roughly categorized into modeling approaches based on sequence homology, protein structure, domain and motif and approaches based on machine learning. These pioneering works are reviewed and discussed below.

2.1. *Homology-based approach*

The homology-based approach is a conventional way for predicting intra-species PPIs. Many studies have also adopted this strategy for predicting host–pathogen PPIs, which are inter-species PPIs. The basic hypothesis of the homology-based approach is that the interaction between a pair of proteins in one species is expected to be conserved in related species.[27] This is a reasonable hypothesis as a pair of homologous proteins descend from the same ancestral pair of interacting proteins and is expected to inherit the structure and function and, thus, interactions of the ancestral proteins. Therefore, the basic procedure of the homology-based approach for intra-species PPI prediction is (i) starting from a known PPI (the template PPI) in some source species, (ii) determining in the target species the homologs (x', y') of the two proteins (x, y) in the template PPIs, and (iii) predicting that the two homologs (x', y') interact in the target species. This approach is generally adapted to the inter-species scenario of host–pathogen PPI prediction by (i) starting from a known PPI (the template PPI) in some source species, (ii) determining in the host a homolog (x') and in the pathogen a homolog (y'), respectively, of the two proteins (x, y) in the template PPI and (iii) predicting that (x', y') interact.

The main advantages of the homology-based approach to host–pathogen PPI prediction are its simplicity and its apparent biological basis. Since the data required for performing the prediction are only the template PPIs and protein sequences, this approach is scalable and can be applied to many different host–pathogen systems. The homology-based approach can be used alone[1–4] or in combination with other methods[5] in predicting host–pathogen PPIs. The investigated host–pathogen systems in past studies include *Homo sapiens–Plasmodium falciparum*,[1,2,5] *H. sapiens–Heliobacter pylori*,[3] *phage T4–Escherichia coli*,[4] *phage lambda–E. coli*,[4] *H. sapiens–E. coli*,[4] *H. sapiens–Salmonella enterica*,[4] *H. sapiens– Yersinia pestis*,[4] etc. The template PPIs used in the prediction can also be very different. The commonly used template PPIs are from DIP,[28] iPfam,[29] MINT,[30] HPRD,[31] Reactome,[32] IntAct,[33] etc.

There is an inherent weakness in the homology-based approach. Basically, in a real biological process, such as infection, the two proteins in a predicted PPI may actually have little opportunity to be present together. Consequently, host–pathogen PPIs predicted solely on the homology basis, without considering other biological properties of the proteins involved, may not be very reliable. Additional information should be used to increase the accuracy of the prediction. For example, extracellular localization and transmembrane regions are used in pruning[4] or constraining the predictions.[3] Also, a pathogen (e.g. *P. falciparum*) may infect different organs at different stages of the pathogen's life cycle. Thus, filtering by tissue-specific gene expression data may also improve prediction reliability.[2] Indeed, recognizing this weakness in the homology-based approach, Wuchty[5] has proposed filtering PPIs predicted by the homology-based approach using a random-forest classifier trained on sequence compositional characteristics of known PPIs, as well as by gene

expression and molecular characteristics. This results in a significantly smaller set of putative host−pathogen PPIs, which are claimed to be of higher quality than the original set of predicted PPIs.

2.2. *Structure-based approach*

When a pair of proteins have structures that are similar to a known interacting pair of proteins, it is reasonable to believe that the former are likely interacting in a way that is structurally similar to the latter. In accordance to this hypothesis, several works have used structural information to identify the similarity between query proteins (i.e. proteins in the pathogen and host) and template PPIs (i.e. known interacting protein pairs) and infer that those host−pathogen protein pairs that match some template PPIs are interacting.

2.2.1. *Comparative modeling*

Prediction by comparative modeling is a representative structure-based approach. For example, in Davis *et al.*,[6] an automated pipeline for large-scale comparative protein structure modeling, MODPIPE, is applied to model the structure of host and pathogen proteins based on their sequences and corresponding template structures. Given the computed model of a protein, the SCOP[34] superfamilies that the protein belongs to are identified. A database of protein structural interfaces, PIBASE, is then scanned. If a SCOP superfamily of a host protein and a SCOP superfamily of a pathogen protein are both involved in the same PIBASE[35] protein structural interface, then the host protein and the pathogen protein are predicted as a putative PPI.

Query proteins that lack structural templates cannot be modeled in the above process. In this case, template interactions in alternative databases (e.g. IntAct) are considered by Davis *et al.*[6] Specifically, a pair of host and pathogen proteins are predicted to interact if at least 50% of each of the two protein sequences are similar to some member proteins of a template complex in IntAct and the joint sequence identity ($\sqrt{Sequence\ Identity1 * Sequence\ Identity2}$) is at least 80%. These predictions, which are conducted without structural information, form a very small portion of the total number of putative PPIs, because of the stringent joint threshold. Each prediction is further followed by a series of assessments and filtering (biological and network filters), which results in a significant reduction of potential host−pathogen PPIs by several order of magnitudes.

2.2.2. *Structural similarity*

Structural similarity can also be analyzed using the Dali database.[36] This strategy has been adopted to predict *H. sapiens*−HIV PPIs,[8] *H. sapiens*−DENV PPIs[7] and *A. aegypti*−DENV PPIs.[7] Dali calculates structural similarity score by comparing the 3D structural coordinates of two PDB entries.[7] To predict the *H. sapiens*−HIV and *H. sapiens*−DENV PPIs, structurally similar pathogen (HIV, DENV) and host

(*H. sapiens*) proteins are first determined using Dali. Then, under the assumption that pathogen proteins having similar structure to host proteins are likely to participate in the similar set of PPIs (*H. sapiens* PPI dataset from HPRD[31]) that those matched host proteins participate in, the pathogen proteins are directly mapped to their high-similarity matches within the host intra-species PPI network to predict the host–pathogen PPIs.[7,8] The same structural similarity prediction method has been applied to identify orthologs between *Drosophila melanogaster* and *Aedes aegypti* and map *D. melanogaster*–DENV PPIs to predict *A. aegypti*–DENV PPIs[7] — the host–pathogen PPIs between DENV and its real insect host. The accuracy of this prediction method depends on the performance of Dali in determining structurally similar pathogen and host proteins. The availability of pathogen and host protein structures and the quality of host intra-species PPI data also have a significant influence on prediction results.

2.3. *Domain and motif interaction-based approach*

Domains are basic building blocks determining the structure and function of proteins and they play specialized role in mediating the interaction of proteins with other molecules.[37] Some studies have proposed predicting host–pathogen PPI based on domain–domain interaction (DDI)[9] and motif–domain interaction.[10]

2.3.1. *Domain–domain interaction-based approach*

Dyer *et al.*[9] predict host–pathogen PPIs in the *H. sapiens–P. falciparum* system by integrating known intra-species PPIs with domain profiles based on an association method (sequence-signature algorithm) proposed by Sprinzak and Margalit.[38] Specifically, domains are first identified by InterProScan[39] in each interacting protein in the intra-species PPIs. Then, the probability $P(d, e)$ that two proteins containing a specific pair of domains (d, e) would interact is estimated for each pair of domains in the Bayesian manner. Finally, given a pair of host–pathogen proteins, their probability of interaction is estimated by a naive combination $(= 1 - \prod_i \prod_j (1 - P(d_i, e_j)))$ of the probabilities from each pair of domains (d_i, e_j) contained in the pair of proteins.[9]

At around the same time, Kim *et al.*[40] predict *H. sapiens–H. pylori* PPIs using the PreDIN[41] and PreSPI[42] algorithms, which are also based on domain information. The domain annotation used in this work is done by InterProScan as well. However, in contrast to Dyer *et al.*,[9] which is based on estimating the probability of an individual pair of domains being associated with protein interactions and naively combining these probabilities, PreDIN and PreSPI directly estimate the probability of domain combination pairs being associated protein interactions.

2.3.2. *Motif–domain interaction-based approach*

Some protein interactions are mediated not by interactions between domains but by interactions between a domain in one protein and a short linear motif (SLiM) in the

other protein.[43,44] As viral pathogens typically have a compact genome, they have few domains. It is reasonable to postulate that their interaction with host proteins are likely to be mediated by Domain−SLiM interactions. For example, since HIV-1 proteins have few domains, Evans *et al.*[10] predicted *H. sapiens*−HIV-1 PPIs based on the interactions between short eukaryotic linear motifs (ELMs) and human protein counter domains (CDs).

Evans *et al.* use the ELM resource[45] to determine ELMs contained in human and HIV-1 proteins and PROSITE[46] to determine domains in human proteins. Then starting from a template human PPI (x, y) where protein x contains a ELM (E) and protein y a counter domain (CD), proteins in HIV-1 that contain the ELM (E) are predicted to form host−pathogen PPIs with the human protein y. Notably, Evans *et al.* point out that the human protein x is expected to compete with these HIV-1 proteins for interacting with y, and that this competition should be considered as another form of host−pathogen interaction.

2.4. *Machine-learning–based approach*

Both supervised[11,12] and semi-supervised[13] learning frameworks have also been used in predicting host−pathogen PPIs. A considerable amount of interacting and non-interacting pairs are usually needed by these machine learning algorithms to produce good classifiers. For example, Tastan *et al.*[11] and Qi *et al.*[13] obtain curated *H. sapiens*−HIV PPIs from the "HIV-1, human protein interaction database",[19] while Dyer *et al.*[12] compile *H. sapiens*−HIV PPIs from other sources including BIND,[47] DIP,[28] IntAct[33] and Reactome.[32] Supervised learning framework has first been attempted using a Random Forest (RF)[11] classifier with 35 selected features, including GO similarity, graph properties of the human interactome, ELM-ligand, gene expression, tissue feature, sequence similarity, post-translational modification similarity to neighbor, HIV-1 protein type, etc. In another work,[12] a Support Vector Machine (SVM) is used with linear kernel and features such as domain profiles, protein sequence k-mers and properties of human proteins in the human interactome.

The performance of supervised learning algorithms is limited by the availability of truly interacting proteins. However, there are a lot of protein pairs that have a known association between themselves which may not be a confirmed direct interaction.[13] In order to exploit the availability of these data, Qi *et al.*[13] try a semi-supervised learning approach.

The semi-supervised approach of Qi *et al.*[13] uses the same training data (collected by Fu *et al.*[19]) as the supervised approach of Tastan *et al.*,[11] who use only physical PPIs with keywords "interact," "bind," etc. for training. However, Qi *et al.*[13] use only a subset of the physical PPIs used by Tastan *et al.*[11] This subset consists of 158 expert-annotated *H. sapiens*−HIV PPIs and is labeled as positive training data. The remaining PPIs from Fu *et al.*[19] are used as "partial positive" training data. This is because Qi *et al.* find that many of the PPIs — even those with keywords "interact," "bind," etc. — are not well agreed by experts.[13] Moreover, only 18 of the 35

attributes used by Tastan *et al.* are used by Qi *et al.* Despite using fewer attributes, the separation of the PPI training data into definite known positive interactions and partial positives helps Qi *et al.* achieve a higher performance than Tastan *et al.*

An important weakness of these approaches based on machine learning is that the features used by them — e.g. the domain profile feature[12] and the HIV-1 protein type feature[11] — are not easy to understand, especially with respect to their biological basis. Another weakness is the limitation of training data. For example, the use of machine learning approaches in the context of host–pathogen PPI prediction has so far been applied in the *H. sapiens*–HIV system because known host–pathogen PPIs are not available in other host–pathogen systems on a sufficiently large scale.

3. Basic Principles of Host–Pathogen Interaction

Some basic principles derived from the analysis of experimentally verified or manually curated host–pathogen PPIs are discussed in this section. These principles either have been reported and confirmed by several works or have high potential to be applied in future works on host–pathogen interactions.

3.1. *Topological properties of targeted host proteins*

Calderwood *et al.*[48] have generated 44 intra-species Epstein-Barr virus (EBV) PPIs and 173 inter-species *H. sapiens*–EBV PPIs using a stringent and systematic two-hybrid system. They observe that the degree (in the human interactome) of human proteins involved in *H. sapiens*–EBV PPIs are significantly higher than randomly selected human proteins. Thus, these targeted human proteins are enriched with hubs (i.e. proteins with high degree in the human interactome).

Moreover, Calderwood *et al.*[48] also report that the minimum number of steps (in terms of PPI edges) between a targeted human protein and a reachable protein in the network is, on average, smaller than that of randomly picked human proteins. Thus the EBV-targeted human proteins have relatively shorter paths to other proteins in the human interactome.[48]

Dyer *et al.*[49] have also analyzed the topological properties of pathogen-targeted host proteins using much larger datasets. The inter-species host–pathogen PPI and intra-species human PPI datasets studied are integrated from primary literature[48] and 7 databases.[28,30–33,47,50] This integrated host–pathogen PPI dataset contains 10,477 experimentally detected and manually curated host–pathogen PPIs, covering 190 pathogens (most of which are viruses), while the integrated human PPI dataset contains 75,457 experimentally verified PPIs.[49] The result reveals that proteins interacting with viral and bacterial pathogen groups tend to have higher degrees (hubs), which confirms one of the observations of Calderwood *et al.*,[48] and higher betweenness centrality (bottlenecks).

Dyer *et al.* also analyze the physical interaction network between human and three bacterial pathogens (*Bacillus anthracis*, *Francisella tularensis* and *Y. pestis*) generated from a modified two-hybrid assay (liquid-format mating).[51] The analyses

show again that pathogens preferentially interact with hubs and bottlenecks in the human interactome.[51] Zhao *et al.*[15] have similarly confirmed that hubs are more likely to be targeted by viruses in studying human−virus PPIs and human signal transduction pathways.

3.2. *Structural properties of host−pathogen PPIs*

Franzosa and Xia[52] report a significant overlap between exogenous (i.e. host−pathogen) and endogenous (i.e. within-host) interfaces of PPIs, suggesting interface mimicry as a possible pathogen strategy to evade immune system detection and to hijack host cellular machinery. The exogenous interactions represent clear cases of horizontal gene transfer between the virus and host.[52] The acquisition of viral protein sequences from hosts are also observed and discussed by Rappoport and Linial.[53]

Comparing with endogenous interfaces, exogenous interfaces tend to be smaller, indicating that the viral genome is under intense selection to reduce its size compared to the host genome.[52] There is a similar observation in another work[53] that viral proteins are noticeably shorter than their corresponding host counterparts, which may result from acquiring only host gene fragment, eliminating internal domain and shortening domain linkers.

Interestingly, Franzosa *et al.*[52] find that virus-targeted interfaces tend to be "date"-like. That is, they are transiently used by different endogenous binding partners at different times and, on average, they utilize more human binding partners than generic endogenous interfaces. This finding is supported by functional enrichment among the mimicked endogenous binding partners for the GO term "Regulation of Biological Process",[52] since proteins involved in biological regulation usually have transient binding with other proteins. This may also partially explain the topological property that targeted host proteins tend to be hubs in the host interactome,[48] because the proteins having date-like interfaces tend to interact with many proteins and appear as hubs in intra-species PPI networks.

Lastly, an analysis of residues involved in exogenous and endogenous interfaces shows that exogenous interfaces are likely to be less conserved then endogenous interfaces.[48]

4. Analysis and Assessment of Host–Pathogen PPIs

Analysis of host−pathogen PPI datasets is essential both for developing better prediction approaches and applying the host−pathogen PPI datasets in the subsequent studies. Assessment and analysis of host−pathogen PPI datasets can be conducted directly using (i) gold standard host−pathogen PPIs or indirectly using (ii) functional information, (iii) localization information, (iv) related experimental data, (v) biological explanation of selected examples, etc.

4.1. *Assessment based on gold standard*

Known truly interacting host–pathogen PPI data (gold standard) are available for a few pathogens. The "HIV-1, Human Protein Interaction database"[19] contains a considerable number of *H. sapiens*–HIV PPIs. A substantial number of host–pathogen PPIs (mainly *H. sapiens*–HIV PPIs) can also be found in other databases including BIND,[47] DIP,[28] IntAct[33] and Reactome.[32] Therefore, in the case of *H. sapiens*–HIV PPIs, a fairly large gold standard dataset is available. For example, the "HIV-1, Human Protein Interaction database"[19] has been used in assessing predictions based on motif–domain interaction.[10] On the other hand, Davis *et al.*[6] have only managed to collect 33 host–pathogen PPIs from the literature to validate their predictions for 10 pathogen species. As another example, Doolittle and Gomez[7] have only managed to collect 3 PPIs from a public database[49] and 20 PPIs from the literature, and only 19 among these collected PPIs are specific to the *H. sapiens*–DENV-2 system that Doolittle and Gomez[7] have made predictions for. Although 9 of these 19 gold standard PPIs are present in the prediction results of Doolittle and Gomez,[7] the assessment has been badly hampered by the small size of the gold standard dataset.

4.2. *Analyses and assessments based on functional information*

4.2.1. *Gene Ontology*

Gene Ontology (GO) terms that are significantly enriched in the host proteins predicted to be targeted by pathogens can be used to evaluate the functional relevance of the predicted host–pathogen PPIs.[6] GO terms specific for human proteins involved in the immune system and for pathogen proteins involved in host–pathogen interactions can also be used to filter putative host–pathogen PPIs.[6]

Several tools can analyze GO term enrichment, including GOstat[54] used by Wuchty,[5] GO::TermFinder[55] used by Davis *et al.*,[6] Ontologizer[56] used by Tanstan *et al.*[11] and DAVID[57] used in many other studies.[7,8,10] Specifically, Wuchty[5] analyzes the GO term enrichment of host proteins in predicted *H. sapiens–P. falciparum* PPIs and derives the 100 most enriched GO terms (in the Biological Process category) of host proteins. He finds that the pathogen may influence important signaling and regulation processes of the host through host–pathogen PPIs.[5] Tastan *et al.*[11] analyze the GO term enrichment of host proteins in predicted host–pathogen PPIs; they find that 31 GO terms in the Molecular Function category (e.g. transcription regulator, ligand-dependent nuclear receptor, MHC class I receptor and protein kinase C activities), 19 GO terms in the Biological Process category (e.g. immune system process and response to stimulus) and 14 GO terms in the Cellular Component category (e.g. membrane-enclosed lumen and plasma membrane) are significantly enriched. Enriched GO terms are identified similarly in several studies[7,8] and, results show consistency with viral infection. Similarly, enriched GO terms have also been analyzed for pathogen groups[49] and Conserved Protein Interaction Modules (CPIM)[51] among *H. sapiens–B. anthracis*, *H. sapiens–F. tularensis* and *H. sapiens– Y. pestis* protein interaction networks.

4.2.2. *Pathway data*

An analysis of host−pathogen PPIs in the context of biological pathways provides a functional overview of the targeted host proteins, illuminates the mechanisms of a pathogen's obstruction on host pathways and serves as an important assessment of predicted host−pathogen PPIs. We first discuss some results derived from an analysis of the known host−pathogen PPIs using pathway data. Then we introduce some assessment strategies of predicted host−pathogen PPIs using pathways.

Balakrishnan *et al.*[58] analyze the PPI dataset from the "HIV-1, Human Protein Interaction database"[19] in the context of human signal transduction in the Pathway Interaction Database (PID)[59] and Reactome.[32] They discover that a majority of human pathways can potentially be targeted by *H. sapiens*−HIV-1 PPIs. However, many alternative paths (starting and ending at the same proteins yet circumventing HIV-1 disrupted intermediate steps) to the HIV-1 targeted paths exist due to human network redundancy; and degradation and downregulation pathways are among the most highly targeted pathways. Singh *et al.*[14] and Zhao *et al.*[15] have also obtained similar results from analyzing the same pathway data: human signal transduction pathways derived from Pathway Interaction Database (PID)[59] and Reactome[32] and virus−host PPI data from VirusMINT.[16] They find that 355 out of 671 pathways are targeted by at least one viral protein. Moreover, the majority of the pathways (268 out of 355) are targeted by more than one viral protein. In these 355 pathways, 413 proteins are targeted by 28 different viruses. Also, 95 of these 413 targeted host proteins are known drug targets.[14,15] However, proteins targeted by different viruses in each pathways are not necessarily the same. Zhao *et al.*[15] further report that centrally located proteins in merged networks of statistically significant pathways are hub proteins and are more frequently targeted by viruses.

Wuchty[5] analyzes both predicted and external (experimentally determined and structurally inferred) *H. sapiens*−*P. falciparum* PPIs using 184 manually curated pathways from PID.[59] He reports that both separate and combined sets of predicted and external PPIs target proteins which have a higher degree and which appear in more pathways.[5] For each pathogen protein, Wuchty[5] identifies pathways enriched with host proteins that are targeted by this pathogen protein using Fisher's exact test. He then constructs a bipartite matrix between pathogen proteins and their corresponding enriched host signaling pathways. Observation of the matrix reveals that the pathogen has many interactions with proteins in the TNF- and NF-kappa B pathways, which indicate the pathogen's obstruction of inflammatory response.[5] To evaluate host−pathogen PPIs predicted by the domain−motif interaction-based approach, KEGG pathway enrichment for HIV-1 proteins (ENV, NEF and TAT) targeted host proteins in the (experimentally verified and computationally predicted) inter-species host−pathogen PPIs are analyzed.[10] The enriched pathways include (i) immune system pathways such as T-cell and B-cell receptor signaling

pathways, apoptosis, focal adhesion and toll-like receptor signaling pathways; (ii) disease pathways such as the colorectal cancer, leukemia and lung cancer pathways; and (iii) signal transduction processes.[10]

4.2.3. *Gene expression data*

Gene expression data are another important functional information source which have been widely used in the filtering, assessment and verification of host–pathogen PPIs. Tissue-specific and infection-related gene expression data are frequently used in host–pathogen studies. A pathogen like *P. falciparum* infects different human organs at different stages of its life cycle. So the expression data of different stages of its life cycle and *H. sapiens* tissue-specific gene expression data can be used simultaneously for pruning putative *H. sapiens–P. falciparum* PPIs.[2,5] For example, *P. falciparum* invades *H. sapiens* liver tissue during the sporozoite stage. The predicted host–pathogen PPIs are thus more likely to be real, if the corresponding human proteins are known to express in liver tissue and the corresponding pathogen proteins are known to express in the sporozoite stage. This filtering strategy has been adopted by several studies.[2,5] For the *H. sapiens–M. tuberculosis* system, human proteins expressed in lung tissue or bronchial epithelial cells and pathogen proteins upregulated in granuloma, pericavity or distal infection sites can be used for filtering purposes.[6] Moreover, pathogen genes involved in *M. tuberculosis* infections[60,61] and human genes involved in *M. tuberculosis*, *L. major* and *T. gondii* infections[62] can be compared with the pathogen and host proteins in predicted *H. sapiens–M. tuberculosis* PPIs as a useful assessment.[6]

4.2.4. *RNA interference data*

RNA interference (RNAi) is a natural process to specifically and selectively inhibit a targeted gene expression. Small interfering RNA (siRNA), short hairpin RNA (shRNA) and bi-functional shRNA are often used to mediate the RNAi effect. Some human proteins, when being silenced by genome-wide RNAi experiments, are found to be nonlethal to human cells but are essential for HIV replication. These human proteins may have high likelihood of interacting with HIV. Therefore, comparing the set of host proteins in predicted host–pathogen PPIs and the set of host proteins identified by RNAi experiments can be used as an assessment. We briefly list some examples below.

Several studies show that knocking down some host proteins by siRNA[63–65] or shRNAs[66] can impair HIV-1 infection or replication. Thus, those host proteins are essential for HIV-1 infection or replication. Therefore, they have higher possibility to interact HIV-1 proteins. This has been used as a filtering criterion[8] and assessment data[10–13] in several studies.

Three works[11–13] based on the machine learning approach for predicting *H. sapiens*–HIV PPIs use an siRNA dataset[64] to assess their prediction results. The assessment is conducted by examining the overlap between the human proteins

targeted by the predicted PPIs and the proteins in the siRNA dataset.[64] Besides, Qi *et al.*[13] also combined four RNAi datasets[63–66] and conducted additional assessment in a similar way.

A five-way comparison has been conducted by Evans *et al.*[10] on five HIV-1 targeted human protein datasets — viz., (i) the human protein dataset targeted by PPIs predicted using the motif–domain interaction-based approach[10]; (ii) human protein dataset targeted by gold standard PPIs from the "HIV-1, Human Protein Interaction database"[19]; and (iii) human protein datasets from three genome-wide RNAi experiments.[63–65] Results show that genome-wide RNAi experiments match each other better than the interaction studies.[10] The matches between protein dataset (i) and the other four protein sets are significant but discrepancies are still observed.[10]

For the *H. sapiens*–DENV system, host protein datasets from two siRNA experiments in DENV infection[67,68] are available. They have also been used to refine *H. sapiens*–DENV PPI prediction result.[7]

4.3. *Pruning based on localization information*

Localization information of pathogen and host proteins may relate to the possibility of their interactions. For extracellular pathogens, their extracellular or secretion proteins may have higher chance of interacting with host surface proteins rather than host nuclear proteins. For intracellular pathogens like viruses, co-localization of host and pathogen proteins may be one of the prerequisites for protein interactions. Several studies use this information to filter prediction results.

4.3.1. *Subcellular localization of host and pathogen proteins*

Since pathogen extracellular and secretion proteins, and proteins with translocational signals are more likely to interact with host extracellular or membrane proteins, such subcellular localization information are often used in pruning of predicted host–pathogen PPIs.[1–5] In connection with this, several tools are used in homology-based approaches[2–4] to predict protein subcellular localization.

4.3.2. *Co-localization of host and pathogen proteins*

As obligate intracellular pathogens, viruses do not have cellular structure or their own metabolism and are solely dependent on the host cell. Therefore, a viral protein and its host protein interaction targets are more likely to be co-localized. Several studies use this basic assumption to assess or filter predicted *H. sapiens*–HIV PPIs[8] and *H. sapiens*-DENV PPIs.[7] Similar information is also used as one of the selected features for classifiers in approaches based on machine learning for predicting *H. sapiens*–HIV PPIs.[11,13] The co-localization information of two proteins can be revealed through their shared GO terms in the Cellular Compartment category.

4.4. *Biological explanation of selected examples*

An analysis of a specific PPI by explaining the underlining biological functions is not an effective assessment of predicted host–pathogen PPIs because such an analysis can cover only a small number of PPIs. However, it may facilitate a better understanding of that putative PPI and therefore promote subsequent experimental verification of that prediction. Explanation of the biological basis of some example PPIs from the whole dataset can be found in many studies.[2–4,6–8,12] Some of the specific examples may have literature or experimental supports, some lack direct literature support but have some indirect supports including structural information, homology to template PPIs, evidence from related experiments (gene expression and RNAi experiment data), etc. Explanation and identification of validated predictions also enhance the impact of the prediction methods; and this approach has been used in many studies.[4,6,12] For example, Dyer *et al.* discuss in detail the predicted *H. sapiens*–HIV PPIs[12] involving the HIV Dependency Factors[64] that have support in the literature.[12] To some extent, explanation of indirect evidence and clues enhances validity of the selected parts of the prediction results.[3,4,6–8] Predicted PPIs both with and without experimental verifications and PPIs involving hypothetical proteins are discussed and explained in Krishnadev and Srinivasan.[4] In another work, Tyagi *et al.*[3] also explain some examples of predicted *H. sapiens–H. pylori* PPIs through the structural point of view and discuss examples of PPIs involving membrane proteins, secreted proteins and hypothetical proteins.

4.5. *Assessment through related experimental data*

Some related experimental data turn out to be useful for assessing the targeted host proteins in host–pathogen PPIs. For example, during budding, host proteins may be incorporated into the virion.[69] Although some host proteins may be taken up by a budding virus accidentally, others are known to play crucial roles in viral life cycle and host–pathogen interaction. A dataset[69] on human protein presents in virion has also been used to filter predicted *H. sapiens*–HIV-1 PPIs.[8]

Qi *et al.*[13] and Tanstan *et al.*[11] use a human protein set hijacked by HIV-1 into its virion[70] to assess their predicted *H. sapiens*–HIV-1 PPIs. Specifically they examine the overlap between targeted human proteins in the predicted PPIs and the 314 human proteins in virion.[70] A large overlap suggests a satisfactory performance of the prediction approach.[11,13]

5. Host–Pathogen Interaction Data Collection and Integration

The rapid progress on the host–pathogen interaction studies is supported by many collection, dissemination, integration, analysis and visualization tools. Host–pathogen interaction databases, can be divided into two categories: (i) collection and curation databases and (ii) integration and analysis databases. There is

no clear dividing line for the two categories. This categorization is mostly for convenience of discussion.

5.1. *Host–pathogen interaction data collection techniques*

Text mining is frequently used for extracting PPI data from literature. This is very useful in facilitating the manual curation of host–pathogen interaction data from publications. For example, VirusMINT[16] relies on a simple text mining approach based on a context-free grammar that identifies sentences containing interaction information to select relevant articles. VirHostNet[17] also uses a text mining approach to prioritize papers for manual curation, where the text mining pipeline is applied to extract keywords related to both virus and experimental procedures.

Moreover, text mining techniques have been applied to specifically extract host–pathogen PPIs from biomedical literature with considerable accuracy.[71] Feature-based and language-based approaches are introduced and compared by Thieu *et al.*[71] Both methods can automatically detect host–pathogen interaction data and extract information about organisms and proteins involved in the interactions.[71] The feature-based method uses SVM trained on features derived from the individual sentences, including names of the organisms and corresponding proteins or genes, keywords describing host–pathogen interaction-specific information, general PPI information, experimental methods and other statistical information.[71] The language-based method uses a link grammar parser combined with semantic patterns derived from training examples.[71]

5.2. *Host–pathogen interaction collection and curation databases*

Host–pathogen interaction collection and curation databases are those dedicated to collect and curate host–pathogen interaction from literature or from experimental data. These databases may have imported some parts of data from other databases but at least contain some data derived from their own collection or curation. Collection and curation databases serve primarily as "data source," and generally provide only simple tools for searching, visualization or analysis. They are often used as the data source for host–pathogen interaction studies or are imported by integration and analysis databases (to be discussed in the next section). In this section we list some representative databases of this category.

PHI-base is a database created to catalog experimentally verified pathogenicity, virulence and effector genes of fungal and Oomycete pathogens.[72] After its update, the PHI-base also covers bacterial pathogens. The pathogens covered by PHI-base infect a wide range of hosts.[18]

The "HIV-1, Human Protein Interaction database" at NCBI aims at cataloging all interactions between HIV-1 and human proteins published in the peer-reviewed literature.[19] Basic search and visualization tools are also provided. It is very popular among the AIDS research community and is well known for its intensive long-term curation effort. The *H. sapiens*–HIV-1 interaction data included in this database

cover both direct and indirect interactions; brief description and PubMed IDs are also provided for each entry. Its *H. sapiens*–HIV-1 PPI data have been used in several studies[10,11,13] and imported as source data by some databases.[16]

The VirusMINT database aims at collecting all interactions between viral and human proteins reported in the literature.[16] It covers more than 110 different viral strains.[16] The curation effort has focused mainly on viruses known to be associated with infectious diseases and oncogenesis in humans.[16] VirusMINT derives its host–virus PPI data from two sources. The first source is from databases of literature-curated PPIs like IntAct,[33] MINT[30] and "HIV-1, Human Protein Interaction database".[19] Host-virus PPI data are uploaded from IntAct and MINT directly without further curation. Only a subset of "HIV-1, Human Protein Interaction database" is imported, which pertains to enzymatic reactions, physical associations and co-localization. The second source is manually curated PPIs from literature; the PPIs are first uploaded to MINT and then re-imported into VirusMINT.[16] The literature curation is facilitated by simple text ming techniques in selecting relevant articles. MINT[30] and VirusMINT are both curated by MINT curators and uploaded first to MINT then to VirusMINT. Much of the PPI data in VirusMINT are the same as in MINT. VirusMINT also provides searching and visualization functions.

VirHostNet (Virus–Host Network) is a management and analysis database of integrated virus–virus, virus–host and host–host interaction networks and their functional annotations.[17] The interaction data are reconstructed from public databases and, for virus–virus and virus–host interactions, are also supplemented by original literature-curated dataset.

A simple text mining strategy has been adopted for prioritizing articles for literature curation. Virus–virus and virus–host interactions data from public databases are also carefully inspected before importing into VirHostNet.[17] Search and visualization functions are supported in this database.

The databases mentioned below are mostly well known for their intra-species PPI datasets. However, their curation and collection have also been extended to inter-species host–pathogen PPIs. IntAct is an open-source, open-data molecular interaction database.[73] Both intra- and inter-species PPI data are collected in this database either from the literature or from direct data depositions. For each PPI entry, a brief description, experimental method and literature citation are included. Several integration databases[16,21,23] import host–pathogen PPI data from IntAct. It is well known for its intensive curation and quality control process. BioGRID (Biological General Repository for Interaction Datasets) archives and disseminates genetic and protein interaction data.[74] BioGRID interaction data are curated from both high-throughput experiments and individual focused studies. Most of the interaction data are intra-species PPIs, but some host–pathogen PPIs are included. DIP (Database of Interacting Proteins) aims to integrate the diverse experimental evidences on PPIs into the database.[28] It is another well-known intra-species PPI integration database. It also collects host–pathogen PPI data. Reactome is a curated, peer-reviewed knowledgebase of biological pathways.[32] It curates both

intra- and inter-species data. Curated host−pathogen PPI data are also available in Reactome.[32] BIND (Biomolecular interaction network database) archives biomolecular interaction, complex and pathway information, and is a major source of curated biomolecular interactions.[47] It has not been maintained for the last few years, until a recent update and conversion of the BIND data to a standard format (Proteomics Standard Initiative-Molecular Interaction 2.5).[75] Its main interaction data are intra-species PPIs, but also contains some host−pathogen PPI data.

5.3. *Host−pathogen interaction integration and analysis databases*

Host−pathogen interaction integration and analysis databases mainly integrate host−pathogen interaction data from other source databases. While they usually do not have their own intensive curation process, some of them provide powerful analysis and visualization functions. The integrated data can be more than just host−pathogen PPI data, like gene expression data related to infection, disease outbreak information, pathogen proteomics data, protein functional data, protein complex data, etc. In this section, representative integration and analysis databases are briefly introduced.

APID (Agile Protein Interaction Data Analyzer) provides an open-access framework where all known experimentally validated PPIs (BIND, BioGRID, DIP, HPRD, IntAct and MINT) are unified in it.[76] iRefIndex[77] provides an index of PPIs from BIND, BioGrid, DIP, HPRD, IntAct, MINT, MPact,[78] MIPS[50] and OPHID.[79] iRefWeb[80] provides a searchable web interface to the iRefIndex. Both APID and iRefIndex (iRefWeb) are general PPI integration databases, unlike the following databases which are dedicated to host−pathogen interaction data integration and analysis. They include host−pathogen PPIs just because their source databases contain some host−pathogen PPI data.

PHIDIAS (Pathogen−Host Interaction Data Integration and Analysis System) includes six components (PGBrowser, Pacodom, BLAST searches, Phinfo, Phigen and Phinet) for searching, comparing and analyzing integrated genome sequences, conserved domains, host−pathogen interaction data and gene expression data related to host−pathogen interactions.[20]

HPIDB is a host−pathogen PPI database which integrates experimental PPIs from several public databases (BIND, REACTOME, MINT, IntAct, PIG).[21] Some of the HPIDB sources may have content overlap with each other, since PIG[23] also integrates data from BIND, REACTOME and MINT. Different from PIG — which only considers one host, *H. sapiens* — HPIDB also takes other hosts into account.

GPS-Prot is an integration and visualization database that currently focuses on *H. sapiens*−HIV interactions.[22] It allows for integration of different HIV interaction data types.[22] Human PPI data are imported from the following six databases, MINT, IntAct, DIP, MIPS, BioGRID and HPRD. *H. sapiens*−HIV PPI data are import from VirusMINT.[16] The GPS-Prot can group proteins into functional modules or

protein complexes, generating intuitive network representations. It allows for the uploading of user-generated data.[22]

RCBPR (Resource Center for Biodefense Proteomics Research) is a bioinformatics framework employing a protein-centric approach to integrate and the collect large and heterogeneous data.[81] It is no longer functional and the collected data have been transferred to the Pathogen Portal (http://www.pathogenportal.org).

PIG (Pathogen Interaction Gateway)[23] is created by integrating host–pathogen PPI data from a number of public resources, including BIND, REACTOME, MINT, MIPS, HPRD, DIP and MvirDB.[82] Now PIG has become part of the PATRIC[25] database but only the bacterial pathogen data in PIG have been merged into PATRIC (which primarily focuses on bacterial pathogens) and can still be accessed at http://www.patricbrc.org/portal/portal/patric/HPITool.

Disease View is a host–pathogen data integration and visualization resource that enables access, analysis and integration of diverse data sources, including host, pathogen, host–pathogen interactions and disease outbreak. It provides a mechanism for infectious-disease–centric data analysis and visualization. The infectious diseases covered by Disease View come with related information like the corresponding pathogen that causes the infectious diseases, the associated pathogen virulence genes and the genetic and chemical evidences for the human genes that are associated with the diseases.[24] It is implemented as a component of PATRIC.[25]

PATRIC (the Pathosystems Resource Integration Center) is a comprehensive genomics-centric relational database for infectious-disease research.[25] Comprehensive bacterial genomics data, associated data relevant to genomic analysis and analysis tools and platforms have been provided in this database. Its resources can be divided into two categories, (i) organisms, genomes and comparative genomics and (ii) recurrent integration of community-derived associated data.

5.4. *Host–pathogen interaction integration and analysis software*

Not only databases but also standalone software tools are available for host–pathogen interaction studies.

Conventional complex network analysis and visualization software platforms like Cytoscape[83] continue to be very popular in host–pathogen interaction studies. Cytoscape has been used for visualization of host–pathogen PPI networks in several works.[9,49] Software that are specifically designed for host–pathogen interaction studies have also been developed. For example, BiologicalNetworks is a system that enables the integration of multiscale data for host–pathogen studies.[26] It can integrate diverse experimental data types, including molecular interactions, phylogenetic classifications, genomic sequences, protein structure information, gene expression, pathway and virulence data for host–pathogen studies.[26] It provides several useful functions, including analyzing subnetworks, building host–pathogen interaction networks, studying individual genes, identifying potential drug targets, adding phylogeography, integrating user data, etc.[26] This system is available through a

standalone Java application (BiologicalNetworks), which provides complex data analysis capabilities, and a web interface (http://flu.sdsc.edu) for quick search of phylogenetic relations among sequenced strains.

6. Discussion

6.1. *Contributions and limitations of current host–pathogen interaction study approaches*

The current host–pathogen interaction studies described in this work are indispensable stepping stones for the future progress in this field. Nevertheless, several limitations are also noticeable.

6.1.1. *Contributions of current host–pathogen interaction studies*

Usually host–pathogen PPIs prediction followed by analyses and assessment would produce enriched datasets which are useful for the experimental testing and verification. This could save a lot of wet lab experimental effort. The prediction and verification approaches discussed in these pioneering works pave the way for future development of host–pathogen interaction studies as they provide insights for improvements and basis for comparison.

6.1.2. *Limitations of current host–pathogen interaction prediction approaches*

It is not uncommon that different prediction approaches yield very different prediction results, even in the same host–pathogen system, as revealed by the comparison among different *H. sapiens*–HIV PPI datasets generated from different prediction approaches.[8]

It has not escaped our notice that some publications repeatedly report almost the same prediction method whose performance and predicted results have not been rigorously assessed. Sometimes even the source data (like template PPI data) are the same, yet only applied to different host–pathogen systems.[2–4] Therefore the contribution of these publications may be relatively limited.

Limited by the current understanding of host–pathogen protein interaction, the prediction approaches may not resemble the real biological scenario. For example, although the approach based on motif–domain interaction[10] has achieved good performance, Evans *et al.* have also mentioned the mismatches between predicted result and gold standard may be caused by the fact that real mechanisms of host–pathogen PPIs are more complicated than the assumption (that host–pathogen PPIs are mediated by ELMs-CDs interactions) in this study.

6.1.3. *Limitations of current host–pathogen interaction verification approaches*

Due to the limitation of current known gold-standard host–pathogen PPI data and limited understanding of the host–pathogen interactions, most current assessments are rather "indirect" approaches.

Some verifications may not have a strong logical or biological basis. For example, Dyer *et al.*[9] assessed predicted *H. sapiens–P. falciparum* PPIs by examining whether the pairs of human proteins predicted to interact with the same pathogen proteins are close to each other in the human PPI network. This assessment through distance in triplets may not have biological or experimental basis. However, based on the observed topological properties discussed in the Sec. 3, calculating whether the human proteins targeted by predicted PPIs have shorter paths to other reachable proteins in the human interactome, would serve as a possible assessment. Dyer *et al.*[9] also analyze the gene expression profile of pathogen protein pairs interacting with the same host proteins; they report that those pathogen protein pairs exhibit correlated gene expression profile, and also the same for host protein pairs interacting with same pathogen proteins. While gene expression profile can be reasonably used in assessing *M. tuberculosis* H37Rv intra-species PPI datasets as done by Zhou and Wong,[84] it may lack biological basis in assessing inter-species host–pathogen PPI dataset through gene expression in the form of triplets as conducted by Dyer *et al.*[9]

Explanation on selected examples of predicted results reflects neither the quality of the whole predicted results nor the performance of prediction approaches. For example, biological explanation for selected examples should not be used as the only assessment of a few predicted results, as what we observed in several studies[1–4] — the qualities of those prediction results are still largely in doubt.

6.2. *Contributions and limitations of current host–pathogen interaction databases*

Current host–pathogen interaction databases contribute a lot to host–pathogen interaction studies in the form of collecting and integrating valuable host–pathogen interactions and providing powerful analysis tools. Yet some possible limitations also exist.

The host–pathogen interaction databases greatly facilitate host–pathogen interaction studies in collecting and integrating valuable interaction and related genomic and experimental data scattered in primary literature. Without these databases, many of the studies described in this review would be impossible or at least would take much longer time and more effort in collecting the source data. Moreover, these databases provide the platforms for accessing and sharing of host–pathogen interaction data, which in turn facilitate research in this field. Many databases not only enable convenient data access and integration of related host and pathogen data but also provide powerful analysis tools which significantly increase the efficiency of host–pathogen interactions analysis.

Some databases lack long-term support and are no longer in function, like RCBPR.[81,85] And there are some information loss in the merging of the one database into another, like PIG,[23] where only its bacterial pathogen data have been moved

into PATRIC. Some databases, although still in operation, lack necessary updates, like ViursMINT[16] and HPIDB.[21]

6.3. *Literature-curated host–pathogen interaction data*

The literature-curated interaction data from the databases discussed above are often used as gold standard in studies on host–pathogen interactions. However, a study[86] on intra-species PPI datasets shows that literature-curated PPI data may not be as accurate as people usually have assumed. Therefore, those manually curated host–pathogen PPI data should be used with caution. For example, the "HIV-1, Human Protein Interaction database" at NCBI[19] has been divided into "positively labeled" and "partially labeled" data in Qi *et al.*,[13] and VirusMINT only imports a portion of the PPI data from it.

6.4. *Future development of host–pathogen interaction studies*

Fundamental progress in host–pathogen interaction studies will be achieved in the future, due to better source data and improved investigation approaches and tools, and these will lead to deeper understanding of host–pathogen interaction.

It is reasonable to expect that high-quality source data will become increasingly available. More genomic and proteomic data will come out. As a result, more accurate orthologs can be identified between less-known pathogens and well-studied organisms. This will enable the application of homology-based approach to many understudied host–pathogen systems. Better annotation of known motifs and counter domains will result in enhanced performance of domain–motif interaction-based prediction approaches.[10] With more protein structures being resolved, the structure-based approach will have higher-quality structural template and with much larger coverage. With more high-resolution Structural Interaction Network (SIN) being provided to analysis, more fundamental interaction mechanisms will come to light. Abundant and accurate functional information — including GO annotation, gene expression, RNA interference and pathway data — will largely improve the performance of current analysis approaches. More reliable PPI data (both intra- and inter-species) will provide sufficient high-quality templates for homology-based approaches and also larger as well as more accurate training and testing data for machine-learning–based approaches. The lack of gold-standard host–pathogen PPI data will also be alleviated in the future. As a result more direct and effective verification approaches will be available for many host–pathogen systems. With better source data from a variety of aspects, the prediction approaches that can integrate different types of data (e.g. machine learning-based approach) into their prediction will have good potential.

More effective host–pathogen interaction prediction algorithms will be proposed in the future. For example, the core algorithm used by Dyer *et al.* is an association method proposed in 2001. However, several other algorithms with enhanced performance are available now, including the association numerical method (ASNM)[87]

in 2004 and the association probabilistic method (APM)[88] in 2006. Using ASNM or APM in predicting host−pathogen PPIs may improve prediction performance. Recently, Itzhaki *et al.*[37] propose the concept of "preferential use of protein domain pairs as interaction mediators" may also introduce new idea to DDI-based prediction algorithms. More accurate prediction and more effective verification approaches on a better understanding of host−pathogen interactions will come out.

All these will help the community to achieve the ultimate goal of better prevention and treatment of infectious diseases.

Acknowledgments

We thank Sriganesh Srihari for the critical reading of this paper. This project was supported in part by an NGS scholarship and a Singapore Ministry of Education Tier-2 grant MOE2009-T2-2-004.

References

1. Lee SA, Chan C, Tsai CH, Lai JM, Wang FS, Kao CY, Huang CY, Ortholog-based protein−protein interaction prediction and its application to inter-species interactions, *BMC Bioinformatics* **9**(Suppl 12):S11, 2008.
2. Krishnadev O, Srinivasan N, A data integration approach to predict host−pathogen protein−protein interactions: Application to recognize protein interactions between human and a malarial parasite, *In Silico Biol* **8**(3):235−250, 2008.
3. Tyagi N, Krishnadev O, Srinivasan N, Prediction of protein−protein interactions between *Helicobacter pylori* and a human host, *Mol BioSyst* **5**(12):1630−1635, 2009.
4. Krishnadev O, Srinivasan N, Prediction of protein−protein interactions between human host and a pathogen and its application to three pathogenic bacteria, *Int J Biol Macromol* **48**:613−619, 2011.
5. Wuchty S, Computational prediction of host-parasite protein interactions between *P. falciparum* and *H. sapiens*, *PLoS ONE* **6**(11):e26960, 2011.
6. Davis FP, Barkan DT, Eswar N, McKerrow JH, Sali A, Host−pathogen protein interactions predicted by comparative modeling, *Protein Sci* **16**(12):2585−2596, 2007.
7. Doolittle JM, Gomez SM, Mapping protein interactions between Dengue virus and its human and insect hosts, *PLoS Negl Trop Disease* **5**(2):e954, 2011.
8. Janet D, Shawn G, Structural similarity-based predictions of protein interactions between HIV-1 and *Homo sapiens*, *Virol J* **7**:82, 2010.
9. Dyer MD, Murali TM, Sobral BW, Computational prediction of host−pathogen protein−protein interactions, *Bioinformatics* **23**(13):i159−i166, 2007.
10. Evans P, Dampier W, Ungar L, Tozeren A, Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs, *BMC Med Genomics* **2**(1):27, 2009.
11. Tastan O, Qi Y, Carbonell JG, Klein-Seetharaman J, Prediction of interactions between HIV-1 and human proteins by information integration, *Pacific Symp Biocomput* **14**:516−527, 2009.
12. Dyer MD, Murali TM, Sobral BW, Supervised learning and prediction of physical interactions between human and HIV proteins, *Infect Genet Evol* **11**(5):917−923, 2011.
13. Qi Y, Tastan O, Carbonell JG, Klein-Seetharaman J, Weston J, Semi-supervised multitask learning for predicting interactions between HIV-1 and human proteins, *Bioinformatics* **26**(18):i645−i652, 2010.

14. Singh I, Tastan O, Klein-Seetharaman J, Comparison of virus interactions with human signal transduction pathways, *Proc First ACM Int Conf Bioinform Comput Biol*, pp. 17−24, 2010.
15. Zhao Z, Xia J, Tastan O, Singh I, Kshirsagar M, Carbonell J, Klein-Seetharaman J, Virus interactions with human signal transduction pathways, *Int J Comput Biol Drug Design* **4** (1):83−105, 2011.
16. Chatr-aryamontri A, Ceol A, Peluso D, Nardozza A, Panni S, Sacco F, Tinti M, Smolyar A, Castagnoli L, Vidal M *et al.*, VirusMINT: A viral protein interaction database, *Nucleic Acids Res* **37**(Suppl 1):D669−D673, 2009.
17. Navratil V, De Chassey B, Meyniel L, Delmotte S, Gautier C, André P, Lotteau V, Rabourdin-Combe C, VirHostNet: A knowledge base for the management and the analysis of proteome-wide virus−host interaction networks, *Nucleic Acids Res* **37**(Suppl 1):D661−D668, 2009.
18. Winnenburg R, Urban M, Beacham A, Baldwin T, Holland S, Lindeberg M, Hansen H, Rawlings C, Hammond-Kosack K, Köhler J, PHI-base update: Additions to the pathogen−host interaction database, *Nucleic Acids Res* **36**(Suppl 1):D572−D576, 2008.
19. Fu W, Sanders-Beer B, Katz K, Maglott D, Pruitt K, Ptak R, Human immunodeficiency virus type 1, human protein interaction database at NCBI, *Nucleic Acids Res* **37**(Suppl 1):D417−D422, 2009.
20. Xiang Z, Tian Y, He Y *et al.*, PHIDIAS: A pathogen-host interaction data integration and analysis system, *Genome Biol* **8**(7):R150, 2007.
21. Ranjit K, Bindu N, HPIDB-a unified resource for host−pathogen interactions, *BMC Bioinformatics* **11**(Suppl 6):S16, 2010.
22. Fahey M, Bennett M, Mahon C, Jäger S, Pache L, Kumar D, Shapiro A, Rao K, Chanda S, Craik C *et al.*, GPS-Prot: A web-based visualization platform for integrating host−pathogen interaction data, *BMC Bioinformatics* **12**(1):298, 2011.
23. Driscoll T, Dyer M, Murali T, Sobral B, PIG — the pathogen interaction gateway, *Nucleic Acids Res* **37**(Suppl 1):D647−D650, 2009.
24. Driscoll T, Gabbard J, Mao C, Dalay O, Shukla M, Freifeld C, Hoen A, Brownstein J, Sobral B, Integration and visualization of host−pathogen data related to infectious diseases, *Bioinformatics* **27**(16):2279−2287, 2011.
25. Gillespie J, Wattam A, Cammer S, Gabbard J, Shukla M, Dalay O, Driscoll T, Hix D, Mane S, Mao C *et al.*, PATRIC: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species, *Infect Immun* **79**(11):4286−4298, 2011.
26. Sergey K, Mayya S, Yulia D, Amarnath G, Animesh R, Julia P, Michael B, BiologicalNetworks-tools enabling the integration of multi-scale data for the host−pathogen studies, *BMC Syst Biol* **5**:7, 2011.
27. Matthews L, Vaglio P, Reboul J, Ge H, Davis B, Garrels J, Vincent S, Vidal M, Identification of potential interaction networks using sequence-based searches for conserved protein−protein interactions or interologs, *Genome Res* **11**(12):2120−2126, 2001.
28. Salwinski L, Miller C, Smith A, Pettit F, Bowie J, Eisenberg D, The database of interacting proteins: 2004 update, *Nucleic Acids Res* **32**(Suppl 1):D449−D451, 2004.
29. Finn R, Marshall M, Bateman A, iPfam: Visualization of protein−protein interactions in PDB at domain and amino acid resolutions, *Bioinformatics* **21**(3):410−412, 2005.
30. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G, MINT: A Molecular INTeraction database, *FEBS Lett* **513**(1):135−140, 2002.
31. Mishra G, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan T *et al.*, Human protein reference database — 2006 update, *Nucleic Acids Res* **34**(Suppl 1):D411−D414, 2006.

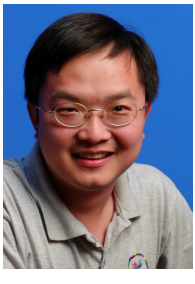
32. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, Bono Bd, Jassal B, Gopinath G, Wu G, Matthews L *et al.*, Reactome: A knowledgebase of biological pathways, *Nucleic Acids Res* **33**(Suppl 1):D428−D432, 2005.
33. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A *et al.*, IntAct: An open source molecular interaction database, *Nucleic Acids Res* **32**(Suppl 1):D452−D455, 2004.
34. Murzin A, Brenner S, Hubbard T, Chothia C *et al.*, Scop: A structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol* **247**(4):536−540, 1995.
35. Davis F, Sali A, PIBASE: A comprehensive database of structurally defined protein interfaces, *Bioinformatics* **21**(9):1901−1907, 2005.
36. Holm L, Kääriäinen S, Rosenström P, Schenkel A, Searching protein structure databases with DaliLite v. 3, *Bioinformatics* **24**(23):2780−2781, 2008.
37. Itzhaki Z, Akiva E, Margalit H, Preferential use of protein domain pairs as interaction mediators: Order and transitivity, *Bioinformatics* **26**(20):2564−2570, 2010.
38. Sprinzak E, Margalit H, Correlated sequence-signatures as markers of protein−protein interaction, *J Mol Biol* **311**(4):681−692, 2001.
39. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R, InterProScan: Protein domains identifier, *Nucleic Acids Res* **33**(Suppl 2):W116−W120, 2005.
40. Kim W, Kim K, Lee E, Marcotte E, Kim H, Suh J, Identification of disease specific protein interactions between the gastric cancer causing pathogen, *H. pylori*, and human hosts using protein network modeling and gene chip analysis, *BioChip J* **1**:179−187, 2007.
41. Kim W, Park J, Suh J *et al.*, Large scale statistical prediction of protein−protein interaction by potentially interacting domain (PID) pair, *Genome Inform Ser* **13**:42−50, 2002.
42. Han D, Kim H, Jang W, Lee S, Suh J, PreSPI: A domain combination based prediction system for protein−protein interaction, *Nucleic Acids Res* **32**(21):6312−6320, 2004.
43. Edwards R, Davey N, Shields D, SLiMFinder: A probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins, *PLoS ONE* **2**(10): e967, 2007.
44. Hugo W, Ng S, Sung W, D-SLIMMER: Domain-SLiM interaction motifs miner for sequence based protein−protein interaction data, *J Proteome Res* **10**(12):5285−5295, 2011.
45. Puntervoll P, Linding R, Gemünd C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin D, Ausiello G, Brannetti B, Costantini A *et al.*, ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins, *Nucleic Acids Res* **31**(13):3625−3630, 2003.
46. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche B, De Castro E, Lachaize C, Langendijk-Genevaux P, Sigrist C, The 20 years of PROSITE, *Nucleic Acids Res* **36**(Suppl 1): D245−D249, 2008.
47. Gilbert D, Biomolecular interaction network database, *Briefings Bioinform* **6**(2):194−198, 2005.
48. Calderwood M, Venkatesan K, Xing L, Chase M, Vazquez A, Holthaus A, Ewence A, Li N, Hirozane-Kishikawa T, Hill D *et al.*, Epstein−Barr virus and virus human protein interaction maps, *Proc Nat Acad Sci USA* **104**(18):7606−7611, 2007.
49. Dyer M, Murali T, Sobral B, The landscape of human proteins interacting with viruses and other pathogens, *PLoS Pathogen* **4**(2):e32, 2008.
50. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes H *et al.*, The MIPS mammalian protein−protein interaction database, *Bioinformatics* **21**(6):832−834, 2005.

51. Dyer M, Neff C, Dufford M, Rivera C, Shattuck D, Bassaganya-Riera J, Murali T, Sobral B, The human-bacterial pathogen protein interaction networks of *Bacillus anthracis, Francisella tularensis* and *Yersinia pestis*, *PLoS ONE* **5**(8):e12089, 2010.
52. Franzosa E, Xia Y, Structural principles within the human-virus protein−protein interaction network, *Proc Nat Acad Sci USA* **108**(26):10538−10543, 2011.
53. Rappoport N, Linial M, Viral proteins acquired from a host converge to simplified domain architectures, *PLoS Comput Biol* **8**(2):e1002364, 2012.
54. Beißbarth T, Speed T, Gostat: Find statistically overrepresented gene ontologies within a group of genes, *Bioinformatics* **20**(9):1464−1465, 2004.
55. Boyle E, Weng S, Gollub J, Jin H, Botstein D, Cherry J, Sherlock G, GO:: TermFinder open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes, *Bioinformatics* **20**(18):3710−3715, 2004
56. Bauer S, Grossmann S, Vingron M, Robinson P, Ontologizer 2.0 a multifunctional tool for go term enrichment analysis and data exploration, *Bioinformatics* **24**(14):1650−1651, 2008.
57. Dennis Jr G, Sherman B, Hosack D, Yang J, Gao W, Lane H, Lempicki R *et al.*, DAVID: Database for annotation, visualization, and integrated discovery, *Genome Biol* **4**(5):P3, 2003.
58. Balakrishnan S, Tastan O, Carbonell J, Klein-Seetharaman J, Alternative paths in HIV-1 targeted human signal transduction pathways, *BMC Genom* **10**(Suppl 3):S30, 2009.
59. Schaefer C, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow K, PID: The pathway interaction database, *Nucleic Acids Res* **37**(Suppl 1):D674−D679, 2009.
60. Sassetti C, Rubin E, Genetic requirements for mycobacterial survival during infection, *Proc Nat Acad Sci USA* **100**(22):12989−12994, 2003.
61. Rachman H, Strong M, Ulrichs T, Grode L, Schuchhardt J, Mollenkopf H, Kosmiadi G, Eisenberg D, Kaufmann S, Unique transcriptome signature of *Mycobacterium tuberculosis* in pulmonary tuberculosis, *Infect Immun* **74**(2):1233−1242, 2006.
62. Chaussabel D, Semnani R, McDowell M, Sacks D, Sher A, Nutman T, Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites, *Blood* **102**(2):672−681, 2003.
63. König R, Zhou Y, Elleder D, Diamond T, Bonamy G, Irelan J, Chiang C, Tu B, De Jesus P, Lilley C *et al.*, Global analysis of host−pathogen interactions that regulate early-stage HIV-1 replication, *Cell* **135**(1):49−60, 2008.
64. Brass A, Dykxhoorn D, Benita Y, Yan N, Engelman A, Xavier R, Lieberman J, Elledge S, Identification of host proteins required for HIV infection through a functional genomic screen, *Science* **319**(5865):921−926, 2008.
65. Zhou H, Xu M, Huang Q, Gates A, Zhang X, Castle J, Stec E, Ferrer M, Strulovici B, Hazuda D *et al.*, Genome-scale RNAi screen for host factors required for HIV replication, *Cell Host Microbe* **4**(5):495−504, 2008.
66. Yeung M, Houzet L, Yedavalli V, Jeang K, A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication, *J Biol Chem* **284**(29):19463−19473, 2009.
67. Sessions O, Barrows N, Souza-Neto J, Robinson T, Hershey C, Rodgers M, Ramirez J, Dimopoulos G, Yang P, Pearson J *et al.*, Discovery of insect and human dengue virus host factors, *Nature* **458**(7241):1047−1050, 2009.
68. Krishnan M, Ng A, Sukumaran B, Gilfoy F, Uchil P, Sultana H, Brass A, Adametz R, Tsui M, Qian F *et al.*, RNA interference screen for human genes associated with West Nile virus infection, *Nature* **455**(7210):242−245, 2008.
69. Chertova E, Chertov O, Coren L, Roser J, Trubey C, Bess Jr J, Sowder II R, Barsov E, Hood B, Fisher R *et al.*, Proteomic and biochemical analysis of purified human

immunodeficiency virus type 1 produced from infected monocyte-derived macrophages, *J Virol* **80**(18):9039−9052, 2006.
70. Ott D, Cellular proteins detected in HIV-1, *Rev Med Virol* **18**(3):159−175, 2008.
71. Thieu T, Joshi S, Warren S, Korkin D, Literature mining of host−pathogen interactions: Comparing feature-based supervised learning and language-based approaches, *Bioinformatics* **28**(6):867−875, 2012.
72. Winnenburg R, Baldwin T, Urban M, Rawlings C, Köhler J, Hammond-Kosack K, PHI-base: A new database for pathogen host interactions, *Nucleic Acids Res* **34**(Suppl 1): D459−D464, 2006.
73. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U *et al.*, The IntAct molecular interaction database in 2012, *Nucleic Acids Res* **40**(D1):D841−D846, 2012.
74. Stark C, Breitkreutz B, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone M, Nixon J, Van Auken K, Wang X, Shi X *et al.*, The BioGRID interaction database: 2011 update, *Nucleic Acids Res* **39**(Suppl 1):D698−D704, 2011.
75. Isserlin R, El-Badrawi R, Bader G, The biomolecular interaction network database in PSI-MI 2.5, *Database: J Biol Databases Curation* **2011**, 2011.
76. Prieto C, De Las Rivas J, APID: Agile Protein Interaction Dataanalyzer, *Nucleic Acids Res* **34**(Suppl 2):W298−W302, 2006.
77. Razick S, Magklaras G, Donaldson I, iRefIndex: A consolidated protein interaction database with provenance, *BMC Bioinformatics* **9**(1):405, 2008.
78. Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes H, Stümpflen V, MPact: The MIPS protein interaction resource on yeast, *Nucleic Acids Res* **34**(Suppl 1): D436−D441, 2006.
79. Brown K, Jurisica I, Online predicted human interaction database, *Bioinformatics* **21**(9):2076−2082, 2005.
80. Turner B, Razick S, Turinsky A, Vlasblom J, Crowdy E, Cho E, Morrison K, Donaldson I, Wodak S, iRefWeb: Interactive analysis of consolidated protein interaction data and their supporting evidence, *Database: J Biol Databases Curation* **2010**, 2010.
81. McGarvey P, Huang H, Mazumder R, Zhang J, Chen Y, Zhang C, Cammer S, Will R, Odle M, Sobral B *et al.*, Systems integration of biodefense omics data for analysis of pathogen-host interactions and identification of potential targets, *PLoS ONE* **4**(9):e7162, 2009.
82. Zhou C, Smith J, Lam M, Zemla A, Dyer M, Slezak T, MvirDB a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications, *Nucleic Acids Res* **35**(Suppl 1):D391−D394, 2007.
83. Smoot M, Ono K, Ruscheinski J, Wang P, Ideker T, Cytoscape 2.8: New features for data integration and network visualization, *Bioinformatics* **27**(3):431−432, 2011.
84. Zhou H, Wong L, Comparative analysis and assessment of *M. tuberculosis* H37Rv protein−protein interaction datasets, *BMC Genomics* **12**(Suppl 3):S20, 2011.
85. Zhang C, Crasta O, Cammer S, Will R, Kenyon R, Sullivan D, Yu Q, Sun W, Jha R, Liu D *et al.*, An emerging cyberinfrastructure for biodefense pathogen and pathogen−host data, *Nucleic Acids Res* **36**(Suppl 1):D884−D891, 2008.
86. Cusick M, Yu H, Smolyar A, Venkatesan K, Carvunis A, Simonis N, Rual J, Borick H, Braun P, Dreze M *et al.*, Literature-curated protein interaction datasets, *Nature Meth* **6**(1):39−46, 2008.
87. Hayashida M, Ueda N, Akutsu T *et al.*, A simple method for inferring strengths of protein−protein interactions, *Genome Inform Ser* **15**(1):56−68, 2004.
88. Chen L, Wu L, Wang Y, Zhang X, Inferring protein interactions from experimental data by association probabilistic method, *Proteins: Struct Funct Bioinform* **62**(4):833−837, 2006.

Hufeng Zhou received both his Bachelor of Arts degree from Huazhong University of Science and Technology and Bachelor of Engineering degree from Huazhong Agricultural University in Wuhan, Hubei, P. R. China, in 2009. Hufeng is currently a Ph.D. candidate at the National University of Singapore. He mainly focuses on the research topics related to host−pathogen interactions, intra-/inter-species protein−protein interaction, pathways integration and analysis, etc.

Jingjing Jin received her Bachelor degree from Sichuan University in Chengdu, Sichuan, China, in 2009. Jingjing is currently a Ph.D. candidate at the National University of Singapore. She mainly focuses on the research topics related to long noncoding RNA and next generation sequencing, etc.



Limsoon Wong is a Professor in the School of Computing and the School of Medicine at the National University of Singapore. He is currently working mostly on knowledge discovery technologies and is especially interested in their application to biomedicine. He serves on the editorial boards of the *Journal of Bioinformatics and Computational Biology* (ICP), *Bioinformatics* (OUP), and *Drug Discovery Today* (Elsevier).