A Review on Host–Pathogen Interactions: Classification and Prediction

Article *in* European Journal of Clinical Microbiology · August 2016

DOI: 10.1007/s10096-016-2716-7

3 authors:

Rishika Sen
Indian Statistical Institute
5 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Losiana Nayak
Indian Statistical Institute
25 PUBLICATIONS **67** CITATIONS
SEE PROFILE

Rajat Kumar De
Indian Statistical Institute
134 PUBLICATIONS **1,140** CITATIONS
SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

A Review on Host-Pathogen Interactions: Classification and Prediction

Rishika Sen · Losiana Nayak · Rajat Kumar De

Received: date / Accepted: date

Abstract The research on host-pathogen interactions is an ever emerging and evolving field. Every other day a new pathogen gets discovered, along with comes the challenge of its prevention and cure. As the intelligent human always vies for prevention which is better than cure, understanding the mechanisms of host-pathogen interactions gets prior importance. There are a whole lot of mechanisms involved from the pathogen as well as the host sides while an interaction happens. It is a *vis-a-vis* fight of the counter genes and proteins from both the sides. Who wins on that depends whether a host gets an infection or not. Moreover, higher level of complexity arises when the pathogens evolve and become resistant to a host's defense mechanisms. Such pathogens pose serious challenges for treatment. The whole human population is in danger of such long-lasting persistent infections. Some of these infections even increase the rate of mortality. Hence there is an immediate emergency to understand how the pathogens interact with their host for successful invasion. It may lead to discovery of appropriate preventive measures, and the development of rational therapeutic measures and medication against such infections and diseases. This review, a state-of-the-art updated scenario of host-pathogen interaction research, has been done by keeping in mind this urgency. It covers the biological and computational aspects of host-pathogen interactions, classification of the methods by which the pathogens interact with their hosts, different machine learning techniques for prediction of host-pathogen interactions and future scopes of this research field.

Keywords Host-Pathogen Interactions · Pathogen Informatics · Machine Learning · *In silico* Prediction · Secretion Systems · Effector Proteins

1 Introduction

The term 'host-pathogen interaction' refers to the ways in which a pathogen (virus, bacteria, prion, fungus and viroid) interacts with its host. Pathogens adapt to the changes, and find alternative ways to survive and infect a host. They are infectious agents which cause diseases in a host body, when the host immune system fails against them. Questions like how the pathogens function, how their entry point into the host is facilitated through the biological barriers and how they survive inside a host that is often under treatment or immunized for the same pathogen, can be answered by exploring host-pathogen interactions. Host-pathogen interactions can be described on the population level (virus infections in a human population), on the organismal level

R. Sen
Machine Intelligence Unit, Indian Statistical Institute
203, Barrackpore Trunk Road, Kolkata 700108, India.
Tel.: +91-9830241122
E-mail: rishikasen_r@isical.ac.in

L. Nayak
Machine Intelligence Unit, Indian Statistical Institute
203, Barrackpore Trunk Road, Kolkata 700108, India.
Tel.: +91-9433477062
E-mail: losiana_t@isical.ac.in

R. K. De
Machine Intelligence Unit, Indian Statistical Institute
203, Barrackpore Trunk Road, Kolkata 700108, India.
Tel.: +91-9433008009
E-mail: rajat@isical.ac.in

(pathogens infecting host), or on the molecular level (pathogen protein binding to a receptor on human cell). However, before stepping into methodological details of host-pathogen interaction processes, a brief glimpse into history of this research field is included here to sum up the how(s) and why(s) of recent advancements of this field.

Some of the earliest research works in the domain of host-pathogen interactions are i) study of host-pathogen interaction in mouse typhoid caused by *Salmonella typhimurium* [146], ii) genetic study of physiology of parasitism of the corn rust pathogen *Puccinia sorghi* [31], iii) a correlation study of α-galactosidase production and host-pathogen interaction between *Phaseolus vulgaris* and *Colletotrichum lindemuthianurn* [42], iv) study of ultrastructural aspects of a host-pathogen relationship of a deuteromycetes fungus, *Pyrenochaeta terrestris* with 2 *Allium cepa* (onion) varieties with the help of electron microscopy [56], v) fine structure study of principal infection procedure during infection of Barley by *Erysiphe graminis* [40], vi) a study on proteins which obstructs the action of the polygalacturonases (polygalaicturonide hydrolases, EC 3.2.1.15) released by the fungal plant pathogens *Fusarium oxysporum*, *Colletotrichum lindemuthianum*, and *Sclerotium rolfsii*. These proteins are extracted from the cell walls of Red Kidney bean hypocotyls, tomato stems and suspension-cultured sycamore cells [1],vii) a study on proteins secreted by plant pathogens which impedes enzymes of the host having the ability to attack the pathogen. The study is conducted on a interaction system of a fungal pathogen (*Colletotrichum lindemuthianum*) and its host, the French bean (*Phaseolus vulgaris*) [2], viii) a study on a single plant protein that efficiently hinders endopolygalacturonases secreted by *Aspergillus niger* and *Colletotrichum lindemuthianum* [46], ix) a molecular basis study to showcase mutation of *Xanthomonas campestris* to overcome resistance in pepper (*Capsicum annuum*) [59], x) a study on stress and immunological response in host-pathogen interactions [90].

Some recent research works have focused on i) the basic notion of virulence and pathogenicity which defines and suggests a classification system for microbial pathogens based on their capacity to cause damage as a consequence of the host's immune response [17], ii) model organisms for host-pathogen interactions, *i.e.*, *C. elegans* [70], *D. melanogaster* [91, 135] and zebrafish [53, 129] among others, iii) molecular cross-talk of host-pathogen interactions where Type III secretion system is mentioned [108], iv) novel studies involving epigenetics[1] [49], metallobiology [11], quantitative temporal viromics[2] [138], heterogeneity in same host tissue [14], and computational systems biology [36] of host-pathogen interactions.

All these investigations indirectly show us the trend of development of the host-pathogen interactions research field. The field has started with sporadic research works of a pathogen and its interaction with a host. The earliest research has been done on host-pathogen interactions with respect to environmental factors, like light, temperature, season, and pathogen/host population among others. Later some organisms, like *C. elegans* and *D. melanogaster* have been found as model organisms to study the pathogen behavior of other complex hosts (human beings) due to their easy body plan, known genome structure and short life cycle. Gradually, certain proteins and then protein clusters have been marked for taking part in host-pathogen interactions. Moreover, definite classification has been found for the mechanism of host-pathogen interactions at the advent of recent developments in imaging and molecular biology techniques.

Moreover, some research works have defined and gave direction to the host-pathogen interactions research field. Discovery of distinct secretion systems [30, 47, 68, 100, 101, 139] has provided the basic background of host-pathogen interaction research. The concerned studies have spanned from genome locus [68] to biochemical and genetic evidence [88]. With discovery of PPI prediction methods [10], the chance of finding host-pathogen protein pairs and their interactions has become more prominent and such studies have given a different direction to the research field. Then methods have been developed for the machine learning based *in silico* prediction of secretion system associated proteins [4]. There are also a couple of newly proposed methods [54, 84] which provide new glimmer of hope to the research field in controlling pathogenesis in a host as described below.

- Secretion systems Type I [139], Type II [30], Type III [47] and Type V [100] have been discovered in 1980s, which have defined the base for host-pathogen interaction research.
- Kuldau *et al.* [68] have predicted 11 ORFs from virB locus in 1990. Based on hydropathy plot they have analyzed that nine of them encode proteins which may interact with membranes and may form a

[1] a procedure through which genotypes give rise to phenotypes during development due to changes in underlying DNA sequence(s), *i.e.*, histone modifications, DNA methylation, DNA silencing via noncoding RNAs and chromatin remodeling proteins.

[2] temporal alterations in host and viral proteins throughout the course of a productive infection

membrane pore or channel to mediate exit of the T-DNA copy. This is the first indirect indication of a distinct secretion system, later known as Type IV Secretion system (T4SS).
- Pukatzki *et al.* have functionally defined T6SS in 2006 [101].
- Mougous *et al.* in 2006 have provided biochemical and genetic evidence that a virulence-associated genetic locus of *P. aeruginosa*, termed as HSI-I, encodes a protein secretion apparatus (T6SS) [88].
- Machine learning based prediction of PPIs have been done by Bock *et al.* in 2001 [10]. They have used Support Vector Machine (SVM) to train and predict interactions based on primary structure and related physicochemical properties. This work has provided a shift in research direction from genes to their protein counter parts and their nature of interaction.
- First ever machine learning based prediction of Type III secretion system associated proteins have been done by Arnold *et al.* in 2009 by analyzing the amino acid composition and secondary structure composition of a few experimentally verified effector proteins at N-terminal [4].
- A few new studies and methods have proposed new avenues of future host-pathogen interaction research, *i.e.*, a new way of studying host-pathogen interaction by dendritic cell subtypes [84] and chemoproteomic profiling of host and pathogen enzymes for finding candidates (proteases) to disrupt pathogenic mechanisms which often have boosted the host's defense mechanisms directly or indirectly [54].

The present review tries encompass the *in silico* prediction of host-pathogen interactions by machine learning and the related aspects. It has been organized into dedicated sections of classification of host-pathogen interactions, availability of host-pathogen interaction data, prediction of host-pathogen interaction domains, image processing based research techniques, and conclusive remarks. There are several substrates and pathways whereby pathogens can invade a host. The human body has its own natural defense mechanism against some of the common pathogens in the form of an immune system that acts against these pathogens. Pathogens have the capability to adhere to host tissues, to evade host defenses, and to invade host cells. However, deeper understanding has revealed that each pathogen has their own variation of these themes [107]. Host-pathogen interactions take place between a host and a pathogen through the protein(s) and gene(s), and by disrupting normal functioning of pathway(s), forming biofilm(s), inhibiting macrophage activity and by other methods. In this review, we have briefly discussed about the various probable factors which directly or indirectly contribute to host-pathogen interactions. Pathogens can either attack a host in gene level by emitting RNA, or they can release proteins which would lead to pathogenicity or they can inhibit the mechanism of macrophage. Some pathogens utilize the components of a host system to survive in the host. These components are called host factors. In a few cases, some factors of a pathogen can initiate the autophagy mechanism which acts in favor of the host. The classification of the host-pathogen interactions is based on traditional pathogen invasion into host.

The review starts with categorization (Figure 1) of pathogens, and makes a comprehensive list of diseases caused by them. The following section discusses classification of host-pathogen Interactions based on different biology based reasoning. Then the widely used *in silico* prediction methods in the domain of host-pathogen interactions are described. Moreover, an extensive list of the online repositories is given. The review concludes with a brief discussion that includes the merits and demerits of this research filed in general, a few scopes for future research and concluding remarks.

2 Classification of Host-Pathogen Interactions

The components of a host-pathogen interaction can be broadly classified into 4 stages, *i.e.*, invasion of host through primary barriers, evasion of host defenses by pathogens, pathogen replication in host and a host's immunological capability to control/eliminate the pathogen. A pathogen can invade a host only after breaching the primary host defenses. Pathogens contain virulence factors which promote and cause disease. The greater the virulence, the more likely the disease will occur. We have classified the host pathogen interactions according to these stages. A summary of the methods discussed in this review has been diagrammatically represented in Figure 2. However, *in silico* prediction methods used for detection of such interactions have been described in the Section 3. The stages mentioned below are overlapping in nature. They do not have a clear boundary between them. The *in silico* prediction methods described later cannot be uniquely associated to only one of the stages. Their applicability spans over many or all the stages of host pathogen interactions.

2.1 Invasion of host through breach of primary barriers

One of the main ways in which pathogens invade the host is via protein secretion. Pathogens, particularly the Gram negative bacteria, which cause pathogenesis in host, consist of secretion systems. These secretion systems release proteins, called effectors, into the body of the host when they come in contact with the host. There are at least six specialized secretion systems in Gram negative bacteria. Type I, Type II, Type III, Type IV, Type V and Type VI are the prominent ones based on their mechanisms of host infection. Details of these mechanisms can be obtained from Costa *et al.* [27]. Numerous secreted proteins are crucial in bacterial pathogenesis. We have described a few of them here, *i.e.*, toxins, urease and multivalent adhesion molecule.

Toxins are substances created by plants and animals that are poisonous to humans. Most toxins that cause problems in humans come from germs such as bacteria. Toxins can be small molecules, peptides, or proteins that are capable of causing disease on contact with or absorption by body tissues interacting with biological macromolecules such as enzymes or cellular receptors. These toxins, once in the body of the host, intervene with the normal functioning of the metabolism of host. Minimized toxin expression in a pathogen have a lesser effect on the stimulation of host's TCR signaling pathway at the time of attack than that with higher toxin expression. It has been observed that viruses interact with different proteins of individual pathways temporally [118]. The molecules that are secreted by gram negative pathogens, lead to damage of the host cells. The vesicle released from the enclosure of the growing bacteria, serves as containers for the proteins and lipids of the Gram negative bacteria. It suggests the importance of vesicle mediated toxin delivery for the onset of infection in the host.

Effectors proteins are secreted by pathogenic bacteria for their entry into host. Effector proteins help a pathogen for invading host tissue, suppressing the host's immune system, and often help the pathogen in its survival. Effector proteins are crucial for virulence. For example, in *Yersinia pestis* (the causative agent of plague), loss of the T3SS has rendered the bacteria completely avirulent [80]. Naive Bayes classifier and support vector machine have already been applied to detect effector proteins of T3SS [4, 136]. More details regarding the methodology is given in the Section 3.

Urease (an enzyme) plays an important role in *Mtb*-host interaction [23]. Urease is present in many species of *mycobacterium*, and its presence/absence is frequently used in the speciation of mycobacteria. Urease has been considered to be a virulence factor for several pathogenic microorganisms. Generation of ammonia by urease of urinary pathogens, such as *P. mirabilis*, have contributed to its pathogenesis due to its toxicity to renal epithelium, participation in complement inactivation and promotion of urinary stone formation [13]. Urease of *H. pylori* alkalinizes the bacterial micro-environment in the stomach and is toxic to stomach epithelium [120]. In the case of *Mtb*, urea is readily available to the bacteria in both its intracellular and extracellular locations within the host.

Multivalent Adhesion Molecule (MAM) is responsible for establishing high affinity binding to host cells during early stages of infection [63]. MAM7 connects to a host via protein-lipid (phosphatidic acid) and protein-protein (fibronectin) interactions. MAM7 has been found on the outer membrane of the gram negative pathogens which contributes to its virulence.

2.2 Evasion of host defenses by pathogens

In order to survive inside the host, the pathogens need to avoid the host defense mechanism. *Mycobacterium tuberculosis (Mtb)* showcases that it actively transcribes a number of genes involved in fortification and evasion from a host system [103]. Assessment of the genome of 58 strains of *Staphylococcus aureus* reveals that all the immune evasive proteins are present in all the strains but not all the surface proteins [81]. Remarkably, 4 strains have surface and immune evasion genes similar to human strain. On the other hand, the putative targets of these proteins vary in different hosts, which propose that these proteins are not crucial for virulence. Signaling for anti-inflammation by glycolipids and host system interaction may be considered as a method of *Mycobacteria* to evade the host or may be playing a vital role in preventing extreme inflammatory response [131].

Pathogens often affect the essential pathways of their hosts with the aim to evade the host defenses. NF-κB family of transcription factors help in the development of the APC (Antigen Presenting Cell) and the lymphocyte [125]. Once the host is compromised, NF-κB pathway gets activated. HIV-1 mostly depends on its host for survival as it has a few genes of its own. An integrated study of HIV-1 and human signal transduction

pathways have been carried out to infer that most of these pathways may get effected by HIV virus during its life cycle [7]. It has assessed and analyzed all possible paths (perturbed and unperturbed) starting from one protein (start point) terminating into another (end point).

Human proteins potentially targeted by EBV (Epstein-Barr virus), tend to be hubs in the human interactome. It is consistent with the hypothesis that hub protein targeting is an effective mechanism for viruses to convert pathways for their use [16]. Bacterial and viral pathogens are more inclined to interact with hub proteins, and the proteins that are central to multiple pathways in the network [38]. Certain cellular mechanisms, like cell cycle regulation and nuclear transport participate in these interactions with a different set of pathogens. A study has identified 3073 human-*B. anthracis*, 1383 human-*F. tularensis* and 4059 human-*Y. pestis* PPIs (Protein-Protein-Interactions) [39]. As suggested by Ranet *et al.* [38], these PPIs have occurred among those hub and bottleneck proteins. The extracellular hydrolytic enzymes, especially the aspartyl proteinases (Saps) secreted by *C. albicans*, are major factors of its pathogenicity [92]. Protein Chaperon 60 and 60.1 have a higher impact on activation of the cytokines than the protein Chaperon 60.2 [75]. In *Staphylococcus aureus*, proteins EsxA and EsxB act as virulent factors to enforce pathogenesis [15]. Mutants that do not secrete these proteins have been observed for failing to enforce strong pathogenesis. Among two closely related families of proteins, PE and PE_PGRS, PE_PGRS of *Mtb* activates a considerable humoral immune response but not PE [29]. Further study suggests that unlike PE, certain PE_PGRS genes are expressed during infection and antibody response. In case of Enterovirus, 71 genes out of 699 get differentially expressed significantly during infection [77]. Lack of the flagella gene in *Salmonella typhimurium* contributes to its virulence. Addition of flagella gene increases the cytotoxicity. However, it does not increase the production of IL-6 (InterLeukin-6) [96].

One of the crucial host defenses is the macrophage. Hence macrophage inhibition is another factor using which the pathogen evades the host immune mechanism. Macrophage activation happens due to multiple components, *i.e.*, gene(s) encoding receptor(s), signal transduction molecule(s), transcription factor(s) and bacterial component(s) that activate toll like receptor(s) (lipopolysacharide, muramyl dipeptide, lipoteichoic acid and heat shock proteins) [94] among others. Pathogens attempt to survive in the host by preventing the macrophages to act on them. It has been found that pathogens disrupt the enzymatic activity in activated macrophages by disrupting the actin filament network [50].

It has been identified that falsatin is an endogenous protease inhibitor of *Plasmodium falciparum*. Analysis of inhibition of normal functionality of macrophages to engulf pathogens and ingest killed parasites due to the functioning of ornithine decarboxylase, has been done by Nairz *et al.* [60]. Due to pathogen specific responses, interleuken-12 production is inhibited for *Mtb*, hence allowing the host to fight against the pathogen. It has been found that 26 to 37 proteins of HIV-1 are associated with MDM (monocyte derived macrophages) derived from HIV [22]. Inhibition by *Mtb* can be avoided with the help of IFN-γ and transfection of LRG-47 [52]. It has been found that *Mtb* residing in macrophage, switches to anaerobic growth [114] to evade host defense for a longer period of time.

The crosstalk of host-pathogen interactions is often governed by miRNAs [48, 111, 112]. The small RNAs, like siRNAs and shRNAs also play a vital role in host-pathogen interactions. Konig *et al.* [62] have studied the association of siRNAs with host-pathogen interactions. They have explored it by combining genome wide siRNA analysis along with the knowledge from human interactome database. Pathogens have Short Linear Motifs (SLiM) that have high similarity with host SLiMs. Motif mimicry is used by pathogens to rewire host signaling pathways by co-opting SLiM-mediated protein interactions to affect the host systems [133].

Pneumolysin (an enzyme) is a key virulence factor [78]. It activates multiple genes and signal transduction pathways in eukaryotes. Cytolytic effect of Pneumolysin contributes to lung injury and neural damage. It sometimes induces apoptosis in neurons and other cells. It can also trigger host mediated apoptosis in macrophages, thus magnifying extermination of pathogens. Pneumolysin has a both way balancing effect on the host.

2.3 Pathogen replication in host

For surviving inside a host, pathogens have multiple ways to facilitate their growth by speedy replication. First of all, they need a few genes and proteins to survive effectively in the host, while a lot more genes and

proteins are required for their survival outside the host. A study on the metabolic network of the pathogen, *Salmonella typhimurium*, has revealed 1083 genes catalyzing 1087 metabolic and transport reactions. This suggests that a minimal set of potent metabolic pathways within *Salmonella typhimurium*, is required for its favorable replication of *Salmonella typhimurium* within the host [104]. Erythrocytic malaria parasite needs proteases for a number of its cellular processes [98] in order to survive in the host.

Pathogens have evolved strategies to promote their survival by performing hijacking of the host cells they infect. Viruses implant their DNA sequence into the normal sequence of these hosts in the hope of their better survival [105] inside the hosts. A genome of the strain of *Mtb*, H37Rv, made up of 4000 genes comprising 4,411,529 base pairs, have a high guanine and cytosine content [24]. In this genome, 194 genes are required for the growth of *Mtb* [110]. A large number of these genes is unique to mycobacteria and its closely related species. It leads to the fact that the mechanism of infection of *Mtb* is different from other pathogenic species.

Some pathogens even respond to more than one micro-environment for their replication and survival. The genes responsible for Snm (secretion in mycobacterium) protein secretion in a mutation of *Mtb*, which is *Mycobacterium smegmatis*, are homologs of their *Mtb* counterpart [26]. It suggests that some strains may have similar secretion mechanism. Four essential gene products (Sm3866, Sm3869, Sm3882c, and Sm3883c) are needed for Snm secretion. *Mtb* exists in various metabolic states. This fact indicates that it may be responsive to more than one micro-environment [45].

The genome of *Mycobacterium tuberculosis* possesses a large family of Ser/Thr protein kinases (STPKs). STPKs have been found to play an important role in cell division and cell envelope biosynthesis [87]. The outer membrane of the bacteria facilitates the interaction between a host and a pathogen [67]. *C. albicans* have the capability to colonize and infect majority of the tissues of human host, which indicates that it can have functionally distinct proteinases (enzymes performing proteolysis) so as to have enough flexibility to multiply and survive in the host.

Sometimes a host itself unknowingly facilitates/inhibits the survival of its pathogens. These facilities are referred to as the host factors. These factors help in pathogen replication, transcription, integration, growth, 198 propagation, pathogen entry, and host-pathogen interactions among others. A set of 295 cellular cofactors (of host) are essential for replication of influenza virus in the early stage [61]. Among these cofactors, 181 are highly significant in host-pathogen interactions, 219 help in efficient influenza virus growth, 23 have role in vital entry and 10 are required for post entry steps of virus replication. Small molecule inhibitors of multiple factors, including vATPase and CAMK2B, go against influenza virus replication. A set of 116 Dengue Virus Host Factors (DVHF) are needed for the propagation of DENV-2 (dengue virus type 2) [115]. Among 82 human homologs of dipteran DVHF, 42 have been identified to be human DVHF. A set of 311 host factors have been found to be responsible for the growth of HIV-1 [148]. Considering HIV dependency factors obtained previously in [12] [148], it is observed that the cardinality of the set of intersection is 311 host factors. Six newly identified host factors are AKT1, PRKAA1, CD97, NEIL3, BMP2k and SERPINB6 [148]. A set of 250 such factors in HIV has been identified [12]. Rab6 and Vps53 play role in viral entry, and TNPO3 is important for viral integration and Med28 for viral transcription. HDF genes show a stronger presence in the immune cell, thus allowing the viruses to evolve in the host cells which perform the life cycle functions needed for them to survive. A set of 213 host factors and 11 HIV encoded proteins have been found responsible for HIV-1 replication [12]. Among them, a few proteins help in regulation of ubiquitin conjugation, DNA damage response, proteolysis and RNA splicing. Forty new factors play a vital role in the process of initiation and/or kinetics of DNA synthesis. Fifteen proteins with different functions have been found to play an significant role in nuclear import or viral DNA integration.

Pathogens, like *M. laprae*, cannot survive independently. Hence, they convert the glial cells of a host into progenitor cells using which it can survive and spread infection inside the host [55]. It alters the genetic structure of the adult Schwann cells to form the progenitor cells. However, it is still unknown how long *M. laprae* can survive in the de-differentiated Schwann cells as they will eventually differentiate back into adult Schwann cells.

Often apoptosis of host factors has been found to be involved in bacterial growth and sustenance inside host [149]. Apoptosis contributes to the processes of host cell deletion method, triggering of inflammation and defense mechanism. Apoptosis by the pathogen *Bordetella pertussis* allows Bordetella to survive in the introductory stages of infection. After the pathogen has successfully colonized the tissue of the host, it stops producing the toxin adenylate cyclase hemolysin.

Biofilm formation plays a major role in host-pathogen interactions. This is a mechanism of pathogens by which they form a biofilm for their survival in the host, often utilizing degraded host proteins *Leucobacter chromiireducens* subsp. solipictus strain TAN 31504 forms biofilm. Exposure to TAN 31504 leads to change in a few innate immunity related genes in *C. elegans* [89]. Esp (a serine protease secreted by *S. epidermidis*) degrades 75 proteins of *Staphylococcus aureus* by proteolytic activity, which include 11 proteins essential for the formation of biofilm [122]. Esp also degrades several human receptor proteins involved in colonization and infection by the pathogen for the benefit of the host.

2.4 A host's immunological capability to control/eliminate the pathogen

In order to prevent occurrence of infection/disease, the host body launches immune response with respect to the pathogenic invasion, *i.e.*, high expression of certain genes [123], autophagy [119, 132], role of dendritic cells [84, 106], glycoconjugates [86, 87] and iron [32, 93] in activation/alteration of host immune system.

Host genes play an important role in its immune response. Mutated β-catenin homolog bar 1 or homeobox gene egl-5 of *C. elegans*, has resulted in defective response and hypersensitivity to *Staphylococcus aureus* [57]. Bar-1 and the fgl-5 genes function parallel to the immune response pathway taken up by *C. elegans*. Over expression of egl-5 resulted in modification of NF-κB dependent TLR2 (Toll-like receptor 2) signaling in epithelial cells suggesting the role played by these two genes in immune defense of a host. Pro-16 in E cadherin is responsible for host specificity towards the human pathogen *Listeria monocytogenes* [73]. E-cadherin of mouse, which is 85% similar to E-cadherin of human, denotes the entry of bacterial pathogen, *Listeria monocytogenes*, by not allowing E-cadherin to interact with bacterial surface protein internalin. If Proline (Pro) in the position 16 of amino acid in human is replaced by Glutamic acid (Glu) then interaction with internalin is disabled. However in mouse, if Glu is substituted by Pro then interaction with internalin is enabled. On *Mtb* interaction with mice, a group of 67 genes in an immuno-competent host has showed a high level of expression than the immuno-deficient host often in 21 days. This shows that 67 genes are responsible for immunity of mice (host) [123].

Autophagy is another mechanism of hosts defense against pathogen. Autophagy can be used in the elimination of *Mtb* [132]. LRG-47 initiates autophagy according to the study carried out by Singh *et al.* [119]. IRGM (Immunity-related GTPase family M protein) also plays role in autophagy and degradation of intracellular bacillary load.

Dendritic cells (DCs) play a vital role in the activation of the immune system on encountering a pathogen [106]. DCs are summoned to the lamina propria of the small intestine after bacterial infection. The number of DCs summoned depends on the pathogenicity of microorganisms confronted. Infection stimulates the release of a variety of soluble factors, including chemokines, which facilitate the summoning of DCs, and cytokines that are strong arbitrators of DC activation. Pathogens, viruses and their components can activate DCs directly. One of the important characteristics of DCs is their ability to migrate. During some infections, this property may have a harmful as well as a favorable side. Relocation of pathogen-laden DCs from the periphery into lymph nodes leads to the activation of T cells. On the other hand, it contributes to the spread of infection within the host.

Glycoconjugates can alter the immune system of human body. Immunomodulatory components of *Mtb* are phosphatidyl-myo-inositol (PMI), lipomannan (LM) and lipoarabinomannan (LAM). Apart from LM and LAM, mannose also contributes to the synthesis of multiple glycosylated proteins and also polymethylated polysaccharides in *Mycobacteria* [86]. These molecules are synthesized by both pathogenic and non-pathogenic species. Many of the genes involved in biosynthesis of these glycoconjugates are important for survival of *Mycobacteria* [109, 110]. Only serine-threonine kinases have been predicted to take part in the regulation process of Mycobacterial glycosyltransferases [3, 87]. The interaction of *Mycobacteria* with the pattern recognition receptors may be an influencing factor for the functioning of the inflammatory signals, hence determining the way in which the immune system reacts [3, 87].

Iron plays an crucial role in the secretion of cytokines and in the activity of the transcription factors, affecting the immune response [32, 93]. Iron homeostasis is controlled by immune cell derived mediators and acute phase proteins. An effective method of host defense is to restrict the supply of iron to the pathogens.

Pathogens have evolved to utilize iron as it is found plenty in the host. The control of iron homeostasis is one of the main issues, as it can be controlled by the host or the pathogen for their benefit.

With such kind of diverse mechanisms involved at each step of pathogen infection, predicting the host-pathogen interactions are extremely crucial. However, prediction of interactions among the huge number of host and pathogen proteins do pose a real-time experimental problem. Hence, many *in silico* prediction methods have been devised to abate such issues. They effectively provide the primary screening of the possible interactions and provide a list of highly probable interactions, which can then be experimentally verified. In the following section, we have listed and described a few of them.

3 Methods for Prediction of Host-Pathogen Interactions

Predictions in the domain of host-pathogen interactions play a vital role in designing rational-therapeutic measures including drugs. Sometimes, experimental procedures can be cumbersome, time-consuming and expensive. Experimenting with all possibilities takes a lot of time. Prediction methods with the help of machine learning can overcome such problems. They can be used to predict the putative data first, which satisfies certain conditions. Then the predicted set can be verified experimentally, which will engage far less time and resources. The respective subsections describe some of the widely used techniques for *in silico* prediction of host-pathogen interactions. One or more of these methods can be used for prediction of genes, proteins, factors and pathways among others of both the host and pathogen. Experimental and data related aspects of these techniques have been covered in Section 2.

3.1 Biological reasoning based prediction of host-pathogen interactions

The most extensively explored way by which a pathogen interacts with the host, is by PPIs. Pathogen proteins interact with host proteins for invading the host. Proteins of a pathogen can affect a host and its environment in multiple ways. They can directly bind with host protein(s) and affect downward cascades of reactions preventing normal function(s) of host. They can even compromise a host's immunological defenses by misguiding and weakening it. They can even utilize the components of a crumbling harsh anaerobic environment of a immune-compromised host. Hence predicting the putative PPIs between a pathogen and its host(s) is of paramount importance. In order to foretell whether a host protein can interact with a pathogen protein or vice-versa, the following categories of methods can be used.

3.1.1 Homology based prediction

An interaction between a pair of proteins in one species is anticipated to be conserved in its related species [79]. Prediction of host-pathogen PPIs in *Homo sapiens* (as host) and *Plasmodium falciparum* (as pathogen) [64] considers interaction templates of human and *P. falciparum* genomic sequences to bring out the probable set on PPIs. Then homology detection algorithm as shown in Figure 3, is applied to these PPIs, to filter out non-homologous ones. The new set thus formed, is made to pass through the filter of stage specific and tissue specific expression data of *P. falciparum* and *Homo sapiens* respectively, and further filtered using the concept of predicted localized data. A study by Lee *et al.* [74] has considered orthologous pair of genes from 18 different species to predict PPIs. Further analyzing them, 81 genes are found to be conserved in all the 18 species, 243 genes are missing in *P. falciparum* but found in the rest of 17 species. Hence, these 81 genes and their related PPIs are probably conserved.

Homology-based approaches to host-pathogen PPI prediction are widely used for their sheer simplicity and biological background support. Since the data needed for implementing the prediction are only the template PPIs and protein sequences, these approaches are adaptable and can be applied to multiple different host-pathogen systems.

Similar is the case of molecular interaction between GBP (Galactose-Binding Protein) and LPS (Gram negative bacterial Lipopolysaccharide). GBP from *Carcinoscorpius rotundicauda* performs as an anti-microbial

defense [76]. Most importantly, GBP shares architectural and functional homology to human proteins. Therefore there is a probability of some human protein and LPS interactions. Moreover, there are 6 Tectonic domains containing LPS binding sites in GBP. GBP acts as a bridge between LPS and CRP (C- Reactive Protein) by indulging in GBP-LPS and GBP-CRP interactions with the aim at forming a stable pathogen recognition molecule. These interactions have indicated that Tectonin domains can differentiate between host and pathogen proteins.

Homology-based approach have their own set of weaknesses. In an infection, two proteins in a predicted PPI may actually have very low probability to be present together. Therefore, host-pathogen PPIs predicted completely on the homology basis, without taking into consideration other biological properties of the proteins involved, may not be very dependable. Further information is needed to increase the accuracy of the prediction. An investigation by Wuchty and Stefan [143] has described filtering of the PPIs predicted by the homology based approach using a random forest classifier. Then the result has been filtered according to expression and molecular characteristics. It has led to a potent subset of proteins that indeed interact.

3.1.2 Structure based prediction

When a pair of proteins have structures that are similar to a known interacting pair of proteins, it is justifiable to believe that the former are likely to interact in a way similar to the latter. Likewise, several investigations have used structural information to recognize the similarity between query proteins (*i.e.*, proteins in the host and pathogen) and template PPIs (*i.e.*, known interacting protein pairs), and conclude that host-pathogen protein pairs, which match some template PPIs, indeed interact. The method is depicted in Figure 4.

A computational method for prediction of PPIs representing host-pathogen interactions has been devised by Davis *et al.* [28]. Their proposed method has first scanned the host and pathogen genome, searched for structural similarity to the already known protein complexes, and then analyzed their probable interactions, using the physical structures of the proteins. The result finally has undergone a filtering by tissue specific expression data of host proteins and stage specific expression data of pathogen proteins, leading to a potent set of proteins that have a high probability to interact.

Mapping of PPIs between the dengue virus, and its human and insect host has been carried out by Doolittle *et al.* [34]. They have also predicted the interactions depending on structural similarity of the host and the pathogen proteins. It has also focused on predictions relevant to stress, unfolded protein response and interferon pathways. Another work by Dolittle *et al.* [33] has predicted PPIs between HIV-1 and *Homo sapiens* based on structural similarity. It has modeled a network of interactions between HIV-I and human proteins. Structurally similar proteins from host and HIV-1, has been retrieved, and from this structurally similar set of proteins, the known interactions has been mapped. The resultant subset has again been screened with factors, like cellular co-localization and RNAi screen to get a more determined set that has higher probability to interact. The result has highlighted on a more potent set of proteins with higher chances of forming PPIs representing the interactions among human and HIV-1.

3.1.3 Domain/motif interaction based prediction

Here the methodology for prediction of host-pathogen PPIs involves integration of known intra-species PPIs with protein domain profiles, and thereby predicting PPIs between a host and a pathogen [37]. For a set of intra-species PPIs, the functional domains are identified for each interacting proteins. For each pair of functional domain, Bayesian statistics is used to compute the possibility of two proteins containing that pair of domain will interact. The method is shown in Figure 5. It has been applied to *Homo sapiens-Plasmodium falciparum* host-pathogen system, and has successfully predicted 516 PPIs. Human proteins anticipated to interact with the same *Plasmodium* protein are close to each other in the human PPI network, and *Plasmodium* pairs predicted to interact with the same human protein are co-expressed in DNA micro-array datasets measured during various stages of the *Plasmodium* life cycle.

Prediction of PPIs, based on motif conserved in HIV-1, has been performed by Evans *et al.* [43] and Bertoletti *et al.* [8]. The similarity between the binding motifs shared by virus and host proteins plays an important part in the crosstalk between virus and host. Similarly, the study by Bertoletti *et al.* [8] has attempted to predict

PPIs based on motif conserved in HIV-1. It has also highlighted the role of chemokines as a factor for liver inflammation.

3.2 Machine learning based predictions of host-pathogen interactions

Machine learning based prediction methods are extensively used for detecting host-pathogen interactions as shown in Table 1. This table lists a few machine learning methods used for prediction of various aspects of host-pathogen interactions, in different species. Moreover, the particular domain knowledge is also included in this table. The sub-area of research in some cases is referred as "Pathogen Informatics". Supervised learning has been used for the prediction of PPIs in the host-pathogen domain by Tastan *et al.* [124]. The work has considered 35 features, including tissue distribution, gene expression profile, gene ontology, graph properties of human interactome, sequence similarity, post-translational modification similarity to neighbor and HIV-1 protein type features among others. Then the authors have selected the top 3 and top 6 features which are of maximum importance to classify the given data set into interacting and non-interacting classes. Random Forest classifier has been used as a tool for supervised learning with these feature set for training and resulting in MAP (Maximum a Posteriori) of 23%. From this computation, it has been concluded that graph and neighbor similarity features contribute to a better classification.

Prediction of proteins secreted by Type III (T3) secretion system has been carried out by Arnold *et al.* [4]. The authors have examined the amino acid composition and the secondary structure of the N-terminal of 100 experimentally verified effector proteins, and used them for identification of T3 secretion signal. They have used Naive Bayes algorithm for classification. The training samples have been grouped depending on how similar they are, and this similarity has been measured by the Smith-Waterman local alignment algorithm. The input feature set has included frequencies of amino acid, amino acid properties and short combinations of them. Finally, the feature selection strategies have been applied to identify the most important feature to do away with computational complexity. In another attempt for prediction, the authors have used derived features from the secondary structure elements. They have used PSIpred software [82] to predict the structure. From the predicted structures, the features of the input vector have been formulated.

In another attempt to predict bacterial Type III secreted (T3S) effectors, a distinct N terminal position-specific amino acid composition feature has been found in more than 50% of T3S proteins [136]. Bi-profile Bayes method has been used in this particular work for feature extraction. Then the entire dataset along with the new feature has been analyzed with a new SVM based classifier. The new classifier has classified T3S and non-T3S proteins successfully.

In order to establish a relation among a host and multiple pathogens, Kshirsagar *et al.* [66] have developed a method taking the similarity in infection initiated by 4 different pathogens in human host. The authors have used machine learning technique in the form of multi-task classification framework. The host-bacteria PPIs have been used as the input to the multi-task classifier, which has then classified the PPIs into interacting and non-interacting classes. Considering the biological hypothesis of similar pathogens targeting the same critical biological processes in a host, the classifier has minimized the empirical error on the training set and favored models that are biased towards the biological hypothesis. A bias term has been incorporated into the classifier in the form of regularizer to overcome it.

A semi supervised multi-task method has been used on *Homo sapiens*-HIV 1 dataset [102] to predict host-pathogen PPIs. The method has involved both supervised and semi-supervised learning. The supervised classifier has worked on labeled PPIs data. The semi-supervised classifier has shared network layers of the supervised classifier and got trained with partially labeled PPIs. This entire framework has been used to improve the recognition of interacting pairs. The supervised classifier has done multi-tasking with a semi-supervised classifier so that weak positive labels could ameliorate the supervised classification.

For prediction of PPIs between *Homo sapiens* and *Plasmodium falciparum*, a random forest classifier has assessed a set of PPIs, and then filtered the result according to expression and molecular characteristics, leading to a subset of proteins which indeed interact among themselves [143]. It has been observed here that the separate sets and a combined set of predicted and experimentally verified interactions have shared similar characteristics. In another investigation, Kshirsagar *et al.* [65] have tried to improve the supervised learning based prediction of PPIs between *Salmonella*-human and *Yersinia*-human. This has been done by replacing

the missing values of the dataset by the values generated by cross species information along with group lasso technique with regularization (obtained 77.6% precision). In order to impute values, localized-nearest neighbor approach (that uses sequence similarity) has been used as the basis to compute locality.

Data mining also forms an integral part of machine learning. Retrieved data about host-pathogen interactions in a few cases reflects information in two different ways, *i.e.*, feature based (SVM) [128] and language based [19]. The investigation by Chaussabel *et al.* [19] has used hierarchical clustering algorithm, by taking the literature available to identify functionally and transcriptionally homologous pair of genes as input. Removal of noise from the PPI databases has been done by removing PPIs that have less probability of taking place. Each such PPI has then been given a score. Then these PPIs have been hierarchically clustered to obtain the PPIs likeliness of occurrence. In this way, it has been found that out of 12122 binary PPIs obtained from BioGRID, 7504 PPIs are less likely to take place.

4 Online Repositories for Host-Pathogen Interactions

Host-pathogen interactions data can be obtained from several databases and repositories. We have summarized some of these repositories in Table 2. Some of these databases are referred purely for their data content, *i.e.*, genome, proteome and metabolic pathway data [137], virus-virus, host-virus and host-host interaction networks [95], PPIs of hosts and pathogens [69], literature based viral-human protein interactions [18], experimentally verified pathogenic, virulence and effector genes of fungal pathogens [140], human signaling and regulatory pathways [113], information on specific biodefense and public health pathogens [121], 3D viral proteins [116], information on invertebrate vectors of human pathogens [71], and a collection of genus specific databases [6] among others. Some of these databases even have integrated in-house tools, *i.e.*, BLAST interface [35] and browser [147] for host-pathogen interactions data analysis. Moreover, we have described some tools [44] used in analysis and visualization of these kinds of data.

PAThosystems Resource Integration Center (PATRIC) [137] includes a relational database, analytical pipelines, and a website that supports querying, browsing, data visualization, and allowing the download of raw and curated data in standard formats. Currently, the database houses complete sequences for viral and bacterial genomes, hence providing an all-inclusive bioinformatics resource for pathogens.

Pathway Interaction Gateway (PIG) provides a text based search and a BLAST interface for searching the host-pathogen PPIs. Each entry in PIG incorporates information on the functional annotations and the domains present in the interacting proteins [35].

VirHostNet (Virus-Host Network) [51, 95] is a public knowledge base specialized in the management and analysis of integrated virus-virus, host-host and virus-host interaction networks coupled with their functional annotations. VirHostNet contains data of virus-host and virus-virus interactions constituting more than 180 distinct viral species. VirHostNet Web interface provides suitable tools which allow effective query and visualization of infected cellular network.

HPIDB (Host-Pathogen Interaction Database) [69] basically contains experimentally confirmed and predicted PPIs of hosts and pathogens.

GPS-Prot [44] is a software tool that permits users to easily create an all-inclusive and integrated HIV-host networks. Its web-based format, which requires no software installation or data downloads, gives it an extra edge over other visualization tools. GPS-Prot enables users to quickly generate networks that amalgamate both genetic and protein-protein interactions between HIV and its human host, into a single representation.

VirusMint [18] contains protein interactions between viral (papilloma viruses, HIV-1, Epstein-Barr, hepatitis B, hepatitis C, herpes and Simian virus 40) and human proteins reported in the literature. VirusMINT presently stores interactions constituting more than 490 unique viral proteins from more than 110 different viral strains.

PHIDIAS (a Pathogen Host Interaction Data Integration and Analysis System) [144] is a database and analysis system to curate, analyze and address different scientific issues in the areas of host-pathogen interactions (PHI, or called host-pathogen interactions or HPI).

MvirDB [147] integrates DNA and protein sequence information from multiple databases. Entries in MvirDB are hyper-linked back to their original sources. A blast tool enables the user to blast against all DNA or protein sequences in MvirDB, and a browser tool enables the user to explore the database to retrieve virulence factor descriptions, sequences and classifications, and to download sequences of interest.

PHI-base [140], a web-accessible database currently catalogs experimentally verified virulence and effector genes from fungal and oomycete pathogens. These pathogens interact with animal, plant and fungi as hosts.

PID [113] is a freely available collection of curated and peer-reviewed pathways composed of human molecular signaling and regulatory events and key cellular processes. PID offers a range of search features to facilitate pathway exploration.

BioHealthBase [121] is a public bioinformatics database and analysis resource for study of specific biodefense and public health pathogens, like *Francisella tularensis*, *Mycobacterium tuberculosis*, *Influenza virus*, *Microsporidia* species and ricin toxin. It serves as a substantial integrated repository of data imported from public databases and data derived from various computational algorithms and information curated from the scientific literature. Its 3D visualization capacity allows researchers to view proteins with their key structural and functional features highlighted.

VPDB (Viral Protein Structural Database) [116] is an interactive database for three dimensional viral proteins. It provides an all-inclusive resource, with an emphasis on the description of derived data from structural biology. At present, VPDB includes viral protein structures from more than 277 viruses with more than 465 virus strains.

VectorBase [71, 72, 85] is a web-accessible data repository storing information about invertebrate vectors of human pathogens. It annotates and maintains vector genomes providing an integrated resource for the research community. It hosts data related to 9 genomes, *i.e.*, mosquitoes (3 *Anopheles gambiae* genome), *Aedes aegypti* and *Culex quinquefasciatus*), body louse (*Pediculus humanus*), tick (*Ixodes scapularis*), tsetse fly (*Glossina morsitans*) kissing bug and (*Rhodnius prolixus*). The data spans across genomic features, expression data, population genetics and ontologies.

EuPathDB [5, 6] is an integrated database covering the eukaryotic pathogens of the genera *Giardia, Cryptosporidium, Neospora, Leishmania, Toxoplasma, Plasmodium, Trypanosoma* and *Trichomonas*. These groups are supported by a taxon-specific database built upon the same infrastructure. EuPathDB portal provides an entry point to all these resources, and the opportunity to leverage orthology for searches across genera.

Similarly, a number of other databases, like PHISTO [127], ViPR [99], HoPaCI-DB [9], VFDB [21] [145] [20], EDWIP [97], AquaPathogen X [41], are available, which help in the host-pathogen interactions domain research.

5 Discussions and Future Scopes

In this section, we discuss multiple faucets of host-pathogen interactions research, the shortcoming of the previously defined methodologies as discussed in Sections 2 and 3 and the future scopes associated with the aforesaid methodologies. It takes both the host and pathogen points of view into account. We discuss the ways in which a pathogen can attack its host, the proteins emitted by a pathogen responsible for perturbing normal functionality of host, the genes responsible for such proteins, silencing and hijacking gene mechanism of pathogens, inhibiting the functions of macrophages, along with genes and proteins needed for their survival inside a host. From the hosts point of view, we also discuss about the factors of pathogen that activates immune response. Salient features of the discussion is given in Table 3.

The genes of multiple strains of an organism have been studied in several investigations [58, 81, 96] to understand the infection mechanism of these strains on the host, and to locate the difference between them. In order to survive in a host, a pathogen can either perform hijacking [105] or it can use the existing environment to survive [12]. The effect of the genes in different strains of a pathogen has been studied. There is still uncertainty in the generalization/specialization of interactions in different strains of pathogens. A study has

suggested that different strains of the same pathogen have different methods of invasion [81]. On the contrary, a counter example has also been provided in [26], which indicates that two strains of *Mycobacterium* have homologous genes required for Snm.

Influenza, DENV-2 and HIV have been in the limelight for identification of the host factors. Other pathogens too need to be taken into account. Inhibition of macrophage is a prospective aspect of research in bioinformatics. The inhibition mechanism needs to be studied in more pathogens apart from the mostly studied ones to find similarity between the inhibition mechanisms among these organisms.

Machine learning based prediction methods have been applied mainly to PPIs. However, protein-ligand interactions and hence prediction of pathways (excluding signal transduction pathways) via machine learning methods have not been attempted much. Different pathogens become drug resistant and form new pathways, and these newly formed pathways can perturb the present host pathways in an unknown way. Similarly, machine learning algorithms in the field of pathway predictions are needed, which would mainly consider protein-ligand binding. Along with reaction dynamics are needed to be known too, as pathways are nothing but chain of reactions. Prediction of Type III secreted bacterial proteins by machine learning techniques is also a challenging task. However, a major drawback in the area of prediction of host-pathogen PPIs, are the unavailability of data sets for different pathogens. Moreover, there is always this lurking issue of biological validation of the predicted PPIs.

Some of the organisms studied for the exploration of host-pathogen PPIs are *Homo sapiens-Plasmodium falciparaum* [37,64,74,143], *Homo sapiens*-Dengue virus [34], *Homo sapiens*-HIV 1 [8,33,43]. However, there are many more host-pathogen pairs waiting in the line for these kinds of studies. In addition, homology-based approaches have their own inherent weaknesses. In real scenario, two proteins in a predicted PPI may actually have little opportunity to be present close enough to interact with each other. Therefore, host-pathogen PPIs predicted entirely on the basis of homology, without considering other biological characteristics of the proteins involved, may not be reliable. Additional information must be used to increase the accuracy of the prediction and make the predictions biologically sound. Keeping this in mind, the study by Wuchty [143] has filtered the predicted PPIs based on homology using gene expression and molecular characteristics. It has led to the formation of a concrete set of PPIs closer to biological scenario. The prediction of PPIs by comparative modeling [28], have very stringent filters leading to the formation of a smaller and robust set of PPIs.

Supervised, unsupervised and semi supervised learning have been mostly used for prediction of host-pathogen PPIs. The organisms for which these predictions have been made are mainly *Homo sapiens-HIV1* [102, 124], *Homo sapiens-Plasmodium falciparum* [143] and *Homo sapiens-Saccharomyces cerevisiae* [25]. Both Tastan *et al.* and Yanjun *et al.* [102,124] have applied their respective algorithms on the same dataset which basically restricts the contribution of the articles. The performance of Random Forest based classifier is negligibly better than the Multi-Layer Perceptron classifier [102]. Some research articles have selected the top 6 and top 3 features among 35 features to predict whether a protein is interacting or not [124]. This is not a novel way of prediction since the interaction between proteins depends on all of its features even if by negligible amount which should not be ignored.

A flaw is often noticed in the choice of a dataset. In a semi-supervised based learning approach to identify PPIs [102], the negative dataset is way extensive than the positive one. The negative (non-interacting) data set has approximately 16000 pairs of proteins while the experimentally verified positive (interacting) dataset has only 158 pairs of protein. Training with such a dataset might lead to a biased classifier and the classifier would be inclined to predict most test pairs as non-interacting. Moreover, the logic used behind selecting non-interacting dataset is based on a random list of pairs of proteins which do not fall into the positive set. It is always a risk, since there is no experimental evidence that the selected negative pairs will not interact at all. There may be several interacting pairs present among the negative set. Another study has been done for predicting proteins secreted by Type III secretion system based only on structural and compositional aspect of the proteins [4]. These studies should include other factors, like expression and molecular characteristics.

One notable thing is that a few attempts have been made on metabolic pathways. For host-pathogen interactions, most of the work has been done with signal transduction pathways. If enzyme(s) from a pathogen is introduced into a host, they get involved with more than one host pathways. There is no tool available which would take a list of protein (enzyme) names and provide the pathway (just one pathway based on these enzymes) based only on those enzymes (at least 90%). Moreover, a pathogen can be associated with more

than one disease. Such diseases, for which a pathogen is responsible, need to be looked into. The scenario becomes more complex, when a host suffers from two or more diseases simultaneously, it implies presence of multiple pathogens responsible for multiple diseases in a host in real time. Such kind of real-time simulation studies are hardly done.

An important aspect that needs to be considered is that some pathogenic proteins prevent the working of macrophage. This is a serious problem in host-pathogen domain. Drugs are needed that would facilitate the working behavior of a macrophage. Drugs are also needed for the prevention of formation of intracytoplasmic vesicle that HIV-1 uses [22] to prevent identification by macrophages. Formation of biofilm [89,122] is another domain that needs to be looked upon. Breaking the biofilm formed by pathogens is indeed recommended to avoid the spread of infection. More attention is needed in this domain, given the rate at which new infectious pathogens are emerging along with their variety of degree of infection.

Hardly any research have been done based on the automated image processing based techniques available for predicting host-pathogen interactions. A study by Mech *et al.* [83] has come up with a technique of a more robust analysis of microscopy images of macrophages that are made to coexist with different *A. fumigatus* strain. Usually the images are manually analyzed, which is time consuming and error prone. The authors used the feature set which includes size, shape, number of cells and cell-cell contacts. By analyzing the images, it has been found that different mutants of *A. fumigatus* have an impact on the ability of the macrophages to adhere and phagocytose the conidia. It has been observed that the rate of phagocytosis is higher in pksP mutant of *A. fumigates*, while it is not the same case in the other strains.

6 Conclusions

In this review, we have covered various aspects of host-pathogen interactions. Interaction of a pathogen with its host(s) is always a unique mechanism. Each one of the pathogenic species has specific mechanism(s) to interact with their host. The different mechanisms of a number of species have been included in this review along with the similarities and similar factors in the attacking mechanism(s) of pathogens. The review has introduced a brief history and introduction of the host-pathogen interactions research field followed by classification of host-pathogen interactions based on gene(s), protein(s), host-factor(s), involved pathway(s) and inhibition mechanism of macrophage(s). It has listed prediction methods used in the host-pathogen interactions domain based on biological reasoning (homology, structure and motif interaction), machine learning (unsupervised, semi-supervised and supervised) and sometimes both the methods. Various data sources used for research in this domain have also been listed. The review concludes with a general discussion of the topic and future scopes followed by a conclusion. The field of host/pathogen interactions is emerging as a crucial area of infectious disease research in the post-genomic era. It is a budding research field where new discoveries are getting announced almost each day throughout the globe. The discovery of dynamics of the host-pathogen interactions will aptly facilitate further development in the field of discovering new drugs and new therapies for different diseases.

Compliance with Ethical Standards

Funding

This work is not funded by any funding agency.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethical approval

This is a review article. Ethical approval is not required as no human subject is involved.

Informed consent

Informed consent is not needed for this work as it is a review article and no human subject is involved.

Author's contributions

RS conceptualized the whole review. She prepared the initial manuscript. LN and RKD gave theoretical input and modified it. RS, LN, and RKD read and corrected the final manuscript.

Acknowledgements

LN acknowledges University Grants Commission, India for a UGC Post-Doctoral Fellowship (No. F.15-1/2013-14/PDFWM-2013-14-GE-ORI-19068(SA-II)).

References

1. Albersheim, P., Anderson, A.J.: Proteins from plant cell walls inhibit polygalacturonases secreted by plant pathogens. Proceedings of the National Academy of Sciences **68**(8), 1815–1819 (1971)
2. Albersheim, P., Valent, B.S.: Host-pathogen interactions VII. Plant pathogens secrete proteins which inhibit enzymes of the host capable of attacking the pathogen. Plant Physiology **53**(5), 684–687 (1974)
3. Alderwick, L.J., Dover, L.G., Seidel, M., Gande, R., Sahm, H., Eggeling, L., Besra, G.S.: Arabinan-deficient mutants of Corynebacterium glutamicum and the consequent flux in decaprenylmonophosphoryl-D-arabinose metabolism. Glycobiology **16**(11), 1073–1081 (2006)
4. Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, H.W., Horn, M., Rattei, T.: Sequence-based prediction of type III secreted proteins. PLoS Pathogen **5**(4), e1000,376 (2009)
5. Aurrecoechea, C., Barreto, A., Brestelli, J., Brunk, B.P., Cade, S., Doherty, R., Fischer, S., Gajria, B., Gao, X., Gingle, A., et al.: EuPathDB: the eukaryotic pathogen database. Nucleic Acids Research **41**(D1), D684–D691 (2013)
6. Aurrecoechea, C., Brestelli, J., Brunk, B.P., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., et al.: EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Research **38**(suppl 1), D415–D419 (2010)
7. Balakrishnan, S., Tastan, O., Carbonell, J., Klein-Seetharaman, J.: Alternative paths in HIV-1 targeted human signal transduction pathways. BMC Genomics **10**(3), 1 (2009)
8. Bertoletti, A., Maini, M.K., Ferrari, C.: The host-pathogen interaction during HBV infection: immunological controversies. Antiviral Therapy **15**(3), 15 (2010)
9. Bleves, S., Dunger, I., Walter, M.C., Frangoulidis, D., Kastenmüller, G., Voulhoux, R., Ruepp, A.: HoPaCI-DB: host-Pseudomonas and Coxiella interaction database. Nucleic Acids Research **42**(D1), D671–D676 (2014)
10. Bock, J.R., Gough, D.A.: Predicting protein– protein interactions from primary structure. Bioinformatics **17**(5), 455– 460 (2001)
11. Botella, H., Stadthagen, G., Lugo-Villarino, G., de Chastellier, C., Neyrolles, O.: Metallobiology of host–pathogen interactions: an intoxicating new insight. Trends in Microbiology **20**(3), 106–112 (2012)
12. Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J., Lieberman, J., Elledge, S.J.: Identification of host proteins required for HIV infection through a functional genomic screen. Science **319**(5865), 921–926 (2008)
13. Breitenbach, J.M., Hausinger, R.P.: Proteus mirabilis urease. Partial purification and inhibition by boric acid and boronic acids. Biochemical Journal **250**(3), 917–920 (1988)
14. Bumann, D.: Heterogeneous host-pathogen encounters: act locally, think globally. Cell Host & Microbe **17**(1), 13–19 (2015)
15. Burts, M.L., Williams, W.A., DeBord, K., Missiakas, D.M.: EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of Staphylococcus aureus infections. Proceedings of the National Academy of Sciences of the United States of America **102**(4), 1169–1174 (2005)
16. Calderwood, M.A., Venkatesan, K., Xing, L., Chase, M.R., Vazquez, A., Holthaus, A.M., Ewence, A.E., Li, N., Hirozane-Kishikawa, T., Hill, D.E., et al.: Epstein–Barr virus and virus human protein interaction maps. Proceedings of the National Academy of Sciences **104**(18), 7606–7611 (2007)
17. Casadevall, A., Pirofski, L.a.: Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. Infection and Immunity **67**(8), 3703–3713 (1999)
18. 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 Research **37**(suppl 1), D669–D673 (2009)
19. Chaussabel, D., Semnani, R.T., McDowell, M.A., Sacks, D., Sher, A., Nutman, T.B.: Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. Blood **102**(2), 672–681 (2003)
20. Chen, L., Xiong, Z., Sun, L., Yang, J., Jin, Q.: VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Research p. gkr989 (2011)
21. Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., Jin, Q.: VFDB: a reference database for bacterial virulence factors. Nucleic Acids Research **33**(suppl 1), D325–D328 (2005)
22. Chertova, E., Chertov, O., Coren, L.V., Roser, J.D., Trubey, C.M., Bess, J.W., Sowder, R.C., Barsov, E., Hood, B.L., Fisher, R.J., et al.: Proteomic and biochemical analysis of purified human immunodeficiency virus type 1 produced from infected monocyte-derived macrophages. Journal of Virology **80**(18), 9039–9052 (2006)
23. Clemens, D.L., Lee, B.Y., Horwitz, M.A.: Purification, characterization, and genetic analysis of Mycobacterium tuberculosis urease, a potentially critical determinant of host-pathogen interaction. Journal of Bacteriology **177**(19), 5644–5652 (1995)
24. Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S., Eiglmeier, K., Gas, S., Barry, C.r., et al.: Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature **393**(6685), 537–544 (1998)

25. Collins, S.R., Kemmeren, P., Zhao, X.C., Greenblatt, J.F., Spencer, F., Holstege, F.C., Weissman, J.S., Krogan, N.J.: Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Molecular & Cellular Proteomics **6**(3), 439–450 (2007)
26. Converse, S.E., Cox, J.S.: A protein secretion pathway critical for Mycobacterium tuberculosis virulence is conserved and functional in Mycobacterium smegmatis. Journal of Bacteriology **187**(4), 1238–1245 (2005)
27. Costa, T.R., Felisberto-Rodrigues, C., Meir, A., Prevost, M.S., Redzej, A., Trokter, M., Waksman, G.: Secretion systems in gram-negative bacteria: structural and mechanistic insights. Nature Reviews Microbiology **13**(6), 343–359 (2015)
28. Davis, F.P., Barkan, D.T., Eswar, N., McKerrow, J.H., Sali, A.: Host–pathogen protein interactions predicted by comparative modeling. Protein Science **16**(12), 2585–2596 (2007)
29. Delogu, G., Brennan, M.J.: Comparative immune response to PE and PE_PGRS antigens of Mycobacterium tuberculosis. Infection and Immunity **69**(9), 5606–5611 (2001)
30. d'Enfert, C., Ryter, A., Pugsley, A.: Cloning and expression in Escherichia coli of the Klebsiella pneumoniae genes for production, surface localization and secretion of the lipoprotein pullulanase. The EMBO journal **6**(11), 3531 (1987)
31. Dickson, J., Syamananda, R., Flangas, A.: The genetic approach to the physiology of parasitism of the corn rust pathogens. American Journal of Botany pp. 614–620 (1959)
32. Doherty, C.P.: Host-pathogen interactions: the role of iron. The Journal of Nutrition **137**(5), 1341–1344 (2007)
33. Doolittle, J.M., Gomez, S.M.: Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. Virology Journal **7**(1), 82 (2010)
34. Doolittle, J.M., Gomez, S.M.: Mapping protein interactions between Dengue virus and its human and insect hosts. PLoS Neglected Tropical Disease **5**(2), e954 (2011)
35. Driscoll, T., Dyer, M.D., Murali, T., Sobral, B.W.: PIG-the pathogen interaction gateway. Nucleic Acids Research **37**(suppl 1), D647–D650 (2009)
36. Durmuş, S., Çakır, T., Özgür, A., Guthke, R.: A review on computational systems biology of pathogen–host interactions. Frontiers in Microbiology **6** (2015)
37. Dyer, M.D., Murali, T., Sobral, B.W.: Computational prediction of host-pathogen protein–protein interactions. Bioinformatics **23**(13), i159–i166 (2007)
38. Dyer, M.D., Murali, T., Sobral, B.W.: The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathogen **4**(2), e32 (2008)
39. Dyer, M.D., Neff, C., Dufford, M., Rivera, C.G., Shattuck, D., Bassaganya-Riera, J., Murali, T., Sobral, B.W.: The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. PloS One **5**(8), e12,089 (2010)
40. Edwards, H., Allen, P., et al.: A fine-structure study of the primary infection process during infection of barley by Erysiphe graminis f. sp. hordei. Phytopathology **60**(10), 1504–1509 (1970)
41. Emmenegger, E., Kentop, E., Thompson, T., Pittam, S., Ryan, A., Keon, D., Carlino, J., Ranson, J., Life, R., Troyer, R., et al.: Development of an aquatic pathogen database (Aquapathogen X) and its utilization in tracking emerging fish virus pathogens in North America. Journal of Fish Diseases **34**(8), 579–587 (2011)
42. English, P.D., Albersheim, P.: Host-pathogen interactions: I. A Correlation between α-galactosidase production and virulence. Plant Physiology **44**(2), 217–224 (1969)
43. Evans, P., Dampier, W., Ungar, L., Tozeren, A.: Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. BMC Medical Genomics **2**(1), 1 (2009)
44. Fahey, M.E., Bennett, M.J., Mahon, C., Jäger, S., Pache, L., Kumar, D., Shapiro, A., Rao, K., Chanda, S.K., Craik, C.S., et al.: GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data. BMC Bioinformatics **12**(1), 298 (2011)
45. Fenhalls, G., Stevens, L., Moses, L., Bezuidenhout, J., Betts, J.C., van Helden, P., Lukey, P.T., Duncan, K.: In situ detection of Mycobacterium tuberculosis transcripts in human lung granulomas reveals differential gene expression in necrotic lesions. Infection and Immunity **70**(11), 6330–6338 (2002)
46. Fisher, M.L., Anderson, A.J., Albersheim, P.: Host-pathogen interactions VI. A single plant protein efficiently inhibits endopolygalacturonases secreted by Colletotrichum lindemuthianum and Aspergillus niger. Plant Physiology **51**(3), 489–491 (1973)
47. Galan, J.E., Curtiss, R.: Cloning and molecular characterization of genes whose products allow Salmonella typhimurium to penetrate tissue culture cells. Proceedings of the National Academy of Sciences **86**(16), 6383–6387 (1989)
48. Ghosh, Z., Mallick, B., Chakrabarti, J.: Cellular versus viral microRNAs in host–virus interaction. Nucleic Acids Research **37**(4), 1035–1048 (2009)
49. Gómez-Díaz, E., Jordà, M., Peinado, M.A., Rivero, A.: Epigenetics of host–pathogen interactions: the road ahead and the road behind. PLoS Pathogen **8**(11), e1003,007 (2012)
50. Guérin, I., de Chastellier, C.: Pathogenic mycobacteria disrupt the macrophage actin filament network. Infection and Immunity **68**(5), 2655–2662 (2000)
51. Guirimand, T., Delmotte, S., Navratil, V.: VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic Acids Research **43**(D1), D583–D587 (2015)
52. Gutierrez, M.G., Master, S.S., Singh, S.B., Taylor, G.A., Colombo, M.I., Deretic, V.: Autophagy is a defense mechanism inhibiting BCG and Mycobacterium tuberculosis survival in infected macrophages. Cell **119**(6), 753–766 (2004)
53. H Meijer, A., P Spaink, H.: Host-pathogen interactions made transparent with the zebrafish model. Current Drug Targets **12**(7), 1000–1017 (2011)
54. Hatzios, S.K., Abel, S., Martell, J., Hubbard, T., Sasabe, J., Munera, D., Clark, L., Bachovchin, D.A., Qadri, F., Ryan, E.T., et al.: Chemoproteomic profiling of host and pathogen enzymes active in cholera. Nature Chemical Biology **12**(4), 268–274 (2016)
55. Hess, S., Rambukkana, A.: Bacterial-induced cell reprogramming to stem cell-like cells: new premise in host–pathogen interactions. Current Opinion in Microbiology **23**, 179–188 (2015)
56. Hess, W.: Ultrastructure of onion roots infected with Pyrenochaeta terrestris, a fungus parasite. American Journal of Botany pp. 832–845 (1969)
57. Irazoqui, J.E., Ng, A., Xavier, R.J., Ausubel, F.M.: Role for β-catenin and HOX transcription factors in Caenorhabditis elegans and mammalian host epithelial-pathogen interactions. Proceedings of the National Academy of Sciences **105**(45), 17,469–17,474 (2008)

58. de Jong, H.K., Parry, C.M., van der Poll, T., Wiersinga, W.J.: Host–pathogen interaction in invasive salmonellosis. PLOS Pathogen **8**(10), e1002,933 (2012)
59. Kearney, B., Ronald, P.C., Dahlbeck, D., Staskawicz, B.J.: Molecular basis for evasion of plant host defence in bacterial spot disease of pepper. Nature **332**(6164), 541–543 (1988)
60. Kierszenbaum, F., Wirth, J.J., McCANN, P.P., Sjoerdsma, A.: Impairment of macrophage function by inhibitors of ornithine decarboxylase activity. Infection and Immunity **55**(10), 2461–2464 (1987)
61. König, R., Stertz, S., Zhou, Y., Inoue, A., Hoffmann, H.H., Bhattacharyya, S., Alamares, J.G., Tscherne, D.M., Ortigoza, M.B., Liang, Y., et al.: Human host factors required for influenza virus replication. Nature **463**(7282), 813–817 (2010)
62. König, R., Zhou, Y., Elleder, D., Diamond, T.L., Bonamy, G.M., Irelan, J.T., Chiang, C.y., Tu, B.P., De Jesus, P.D., Lilley, C.E., et al.: Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. Cell **135**(1), 49–60 (2008)
63. Krachler, A.M., Ham, H., Orth, K.: Outer membrane adhesion factor multivalent adhesion molecule 7 initiates host cell binding during infection by gram-negative pathogens. Proceedings of the National Academy of Sciences **108**(28), 11,614–11,619 (2011)
64. 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 Biology **8**(3, 4), 235–250 (2008)
65. Kshirsagar, M., Carbonell, J., Klein-Seetharaman, J.: Techniques to cope with missing data in host–pathogen protein interaction prediction. Bioinformatics **28**(18), i466–i472 (2012)
66. Kshirsagar, M., Carbonell, J., Klein-Seetharaman, J.: Multitask learning for host–pathogen protein interactions. Bioinformatics **29**(13), i217–i226 (2013)
67. Kuehn, M.J., Kesty, N.C.: Bacterial outer membrane vesicles and the host–pathogen interaction. Genes & Development **19**(22), 2645–2655 (2005)
68. Kuldau, G.A., De Vos, G., Owen, J., McCaffrey, G., Zambryski, P.: The virB operon of Agrobacterium tumefaciens pTiC58 encodes 11 open reading frames. Molecular and General Genetics MGG **221**(2), 256–266 (1990)
69. Kumar, R., Nanduri, B.: HPIDB-a unified resource for host-pathogen interactions. BMC Bioinformatics **11**(6), 1 (2010)
70. Kurz, C.L., Ewbank, J.J.: Caenorhabditis elegans for the study of host–pathogen interactions. Trends in Microbiology **8**(3), 142–144 (2000)
71. Lawson, D., Arensburger, P., Atkinson, P., Besansky, N.J., Bruggner, R.V., Butler, R., Campbell, K.S., Christophides, G.K., Christley, S., Dialynas, E., et al.: VectorBase: a home for invertebrate vectors of human pathogens. Nucleic Acids Research **35**(suppl 1), D503–D505 (2007)
72. Lawson, D., Arensburger, P., Atkinson, P., Besansky, N.J., Bruggner, R.V., Butler, R., Campbell, K.S., Christophides, G.K., Christley, S., Dialynas, E., et al.: VectorBase: a data resource for invertebrate vector genomics. Nucleic Acids Research **37**(suppl 1), D583–D587 (2009)
73. Lecuit, M., Dramsi, S., Gottardi, C., Fedor-Chaiken, M., Gumbiner, B., Cossart, P.: A single amino acid in E-cadherin responsible for host specificity towards the human pathogen Listeria monocytogenes. The EMBO journal **18**(14), 3956–3963 (1999)
74. Lee, S.A., Chan, C.h., Tsai, C.H., Lai, J.M., Wang, F.S., Kao, C.Y., Huang, C.Y.F.: Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. BMC Bioinformatics **9**(12), 1 (2008)
75. Lewthwaite, J.C., Coates, A.R., Tormay, P., Singh, M., Mascagni, P., Poole, S., Roberts, M., Sharp, L., Henderson, B.: Mycobacterium tuberculosis Chaperonin 60.1 Is a More Potent Cytokine Stimulator than Chaperonin 60.2 (Hsp 65) and Contains a CD14-Binding Domain. Infection and Immunity **69**(12), 7349–7355 (2001)
76. Low, D.H.P., Frecer, V., Le Saux, A., Srinivasan, G.A., Ho, B., Chen, J., Ding, J.L.: Molecular interfaces of the galactose-binding protein Tectonin domains in host-pathogen interaction. Journal of Biological Chemistry **285**(13), 9898–9907 (2010)
77. Lui, Y.L.E., Tan, T.L., Timms, P., Hafner, L.M., Tan, K.H., Tan, E.L.: Elucidating the host–pathogen interaction between human colorectal cells and invading Enterovirus 71 using transcriptomics profiling. FEBS open bio **4**(1), 426–431 (2014)
78. Marriott, H.M., Mitchell, T.J., Dockrell, D.H.: Pneumolysin: a double-edged sword during the host-pathogen interaction. Current Molecular Medicine **8**(6), 497–509 (2008)
79. Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., Vidal, M.: Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Research **11**(12), 2120–2126 (2001)
80. Mattoo, S., Lee, Y.M., Dixon, J.E.: Interactions of bacterial effector proteins with host proteins. Current Opinion in Immunology **19**(4), 392–401 (2007)
81. McCarthy, A.J., Lindsay, J.A.: Genetic variation in Staphylococcus aureus surface and immune evasion genes is lineage associated: implications for vaccine design and host-pathogen interactions. BMC Microbiology **10**(1), 1 (2010)
82. McGuffin, L.J., Bryson, K., Jones, D.T.: The PSIPRED protein structure prediction server. Bioinformatics **16**(4), 404–405 (2000)
83. Mech, F., Thywißen, A., Guthke, R., Brakhage, A.A., Figge, M.T.: Automated image analysis of the host-pathogen interaction between phagocytes and Aspergillus fumigatus. PloS One **6**(5), e19,591 (2011)
84. Mege, J.L.: Dendritic cell subtypes: a new way to study host-pathogen interaction. Virulence **7**(1), 5–6 (2016)
85. Megy, K., Emrich, S.J., Lawson, D., Campbell, D., Dialynas, E., Hughes, D.S., Koscielny, G., Louis, C., MacCallum, R.M., Redmond, S.N., et al.: VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. Nucleic Acids Research **40**(D1), D729–D734 (2012)
86. Mishra, A.K., Driessen, N.N., Appelmelk, B.J., Besra, G.S.: Lipoarabinomannan and related glycoconjugates: structure, biogenesis and role in Mycobacterium tuberculosis physiology and host–pathogen interaction. FEMS Microbiology Reviews **35**(6), 1126–1157 (2011)
87. Molle, V., Kremer, L.: Division and cell envelope regulation by Ser/Thr phosphorylation: Mycobacterium shows the way. Molecular Microbiology **75**(5), 1064–1077 (2010)
88. Mougous, J.D., Cuff, M.E., Raunser, S., Shen, A., Zhou, M., Gifford, C.A., Goodman, A.L., Joachimiak, G., Ordoñez, C.L., Lory, S., et al.: A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus. Science **312**(5779), 1526–1530 (2006)
89. Muir, R.E., Tan, M.W.: Virulence of Leucobacter chromiireducens subsp. solipictus to Caenorhabditis elegans: characterization of a novel host-pathogen interaction. Applied and Environmental Microbiology **74**(13), 4185–4198 (2008)
90. Murray, P.J., Young, R.A.: Stress and immunological recognition in host-pathogen interactions. Journal of Bacteriology **174**(13), 4193 (1992)

91. Mylonakis, E., Aballay, A.: Worms and flies as genetically tractable animal models to study host-pathogen interactions. Infection and Immunity **73**(7), 3833–3841 (2005)
92. Naglik, J., Albrecht, A., Bader, O., Hube, B.: Candida albicans proteinases and host/pathogen interactions. Cellular Microbiology **6**(10), 915–926 (2004)
93. Nairz, M., Schroll, A., Sonnweber, T., Weiss, G.: The struggle for iron–a metal at the host–pathogen interface. Cellular Microbiology **12**(12), 1691–1702 (2010)
94. Nau, G.J., Richmond, J.F., Schlesinger, A., Jennings, E.G., Lander, E.S., Young, R.A.: Human macrophage activation programs induced by bacterial pathogens. Proceedings of the National Academy of Sciences **99**(3), 1503–1508 (2002)
95. 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 Research **37**(suppl 1), D661–D668 (2009)
96. Olsen, J.E., Hoegh-Andersen, K.H., Casadesús, J., Rosenkranzt, J., Chadfield, M.S., Thomsen, L.E.: The role of flagella and chemotaxis genes in host pathogen interaction of the host adapted Salmonella enterica serovar Dublin compared to the broad host range serovar S. Typhimurium. BMC Microbiology **13**(1), 1 (2013)
97. Onstad, D.W.: Ecological Database of the World's Insect Pathogens (edwip) (1997)
98. Pandey, K.C., Singh, N., Arastu-Kapur, S., Bogyo, M., Rosenthal, P.J.: Falstatin, a cysteine protease inhibitor of Plasmodium falciparum, facilitates erythrocyte invasion. PLoS Pathogen **2**(11), e117 (2006)
99. Pickett, B.E., Sadat, E.L., Zhang, Y., Noronha, J.M., Squires, R.B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., et al.: ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Research **40**(D1), D593–D598 (2012)
100. Pohlner, J., Halter, R., Beyreuther, K., Meyer, T.F.: Gene structure and extracellular secretion of Neisseria gonorrhoeae IgA protease. Nature **325**(6103), 458–462 (1986)
101. Pukatzki, S., Ma, A.T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W.C., Heidelberg, J.F., Mekalanos, J.J.: Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. Proceedings of the National Academy of Sciences **103**(5), 1528–1533 (2006)
102. Qi, Y., Tastan, O., Carbonell, J.G., Klein-Seetharaman, J., Weston, J.: Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. Bioinformatics **26**(18), i645–i652 (2010)
103. Rachman, H., Strong, M., Ulrichs, T., Grode, L., Schuchhardt, J., Mollenkopf, H., Kosmiadi, G.A., Eisenberg, D., Kaufmann, S.H.: Unique transcriptome signature of Mycobacterium tuberculosis in pulmonary tuberculosis. Infection and Immunity **74**(2), 1233–1242 (2006)
104. Raghunathan, A., Reed, J., Shin, S., Palsson, B., Daefler, S.: Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction. BMC Systems Biology **3**(1), 38 (2009)
105. Rappoport, N., Linial, M.: Viral proteins acquired from a host converge to simplified domain architectures. PLoS Computational Biology **8**(2), e1002,364 (2012)
106. Rescigno, M., Borrow, P.: The host-pathogen interaction: new themes from dendritic cell biology. Cell **106**(3), 267–270 (2001)
107. Rupp, S., Sohn, K.: Host-pathogen interactions: methods and protocols. Humana Press (2009)
108. Sansonetti, P.: Host–pathogen interactions: the seduction of molecular cross talk. Gut **50**(suppl 3), iii2–iii8 (2002)
109. Sassetti, C.M., Boyd, D.H., Rubin, E.J.: Genes required for mycobacterial growth defined by high density mutagenesis. Molecular Microbiology **48**(1), 77–84 (2003)
110. Sassetti, C.M., Rubin, E.J.: Genetic requirements for mycobacterial survival during infection. Proceedings of the National Academy of Sciences **100**(22), 12,989–12,994 (2003)
111. Scaria, V., Hariharan, M., Maiti, S., Pillai, B., Brahmachari, S.K.: Host-virus interaction: a new role for microRNAs. Retrovirology **3**(1), 68 (2006)
112. Scaria, V., Hariharan, M., Pillai, B., Maiti, S., Brahmachari, S.K.: Host–virus genome interactions: macro roles for microRNAs. Cellular Microbiology **9**(12), 2784–2794 (2007)
113. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H.: PID: the pathway interaction database. Nucleic Acids Research **37**(suppl 1), D674–D679 (2009)
114. Schnappinger, D., Ehrt, S., Voskuil, M.I., Liu, Y., Mangan, J.A., Monahan, I.M., Dolganov, G., Efron, B., Butcher, P.D., Nathan, C., et al.: Transcriptional adaptation of Mycobacterium tuberculosis within macrophages insights into the phagosomal environment. The Journal of Experimental Medicine **198**(5), 693–704 (2003)
115. Sessions, O.M., Barrows, N.J., Souza-Neto, J.A., Robinson, T.J., Hershey, C.L., Rodgers, M.A., Ramirez, J.L., Dimopoulos, G., Yang, P.L., Pearson, J.L., et al.: Discovery of insect and human dengue virus host factors. Nature **458**(7241), 1047–1050 (2009)
116. Sharma, O.P., Jadhav, A., Hussain, A., Kumar, M.S.: VPDB: Viral protein structural database. Bioinformation **6**(8), 324 (2011)
117. Simeone, R., Bottai, D., Brosch, R.: Esx/type vii secretion systems and their role in host–pathogen interaction. Current Opinion in Microbiology **12**(1), 4–10 (2009)
118. Singh, I., Tastan, O., Klein-Seetharaman, J.: Comparison of virus interactions with human signal transduction pathways. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, pp. 17–24. ACM (2010)
119. Singh, S.B., Davis, A.S., Taylor, G.A., Deretic, V.: Human IRGM induces autophagy to eliminate intracellular mycobacteria. Science **313**(5792), 1438–1441 (2006)
120. Smoot, D., Mobley, H., Chippendale, G., Lewison, J., Resau, J.: Helicobacter pylori urease activity is toxic to human gastric epithelial cells. Infection and Immunity **58**(6), 1992–1994 (1990)
121. Squires, B., Macken, C., Garcia-Sastre, A., Godbole, S., Noronha, J., Hunt, V., Chang, R., Larsen, C.N., Klem, E., Biersack, K., et al.: BioHealthBase: informatics support in the elucidation of influenza virus host–pathogen interactions and virulence. Nucleic Acids Research **36**(suppl 1), D497–D503 (2008)
122. Sugimoto, S., Iwamoto, T., Takada, K., Okuda, K.i., Tajima, A., Iwase, T., Mizunoe, Y.: Staphylococcus epidermidis Esp degrades specific proteins associated with Staphylococcus aureus biofilm formation and host-pathogen interaction. Journal of Bacteriology **195**(8), 1645–1655 (2013)
123. Talaat, A.M., Lyons, R., Howard, S.T., Johnston, S.A.: The temporal expression profile of Mycobacterium tuberculosis infection in mice. Proceedings of the National Academy of Sciences of the United States of America **101**(13), 4602–4607 (2004)

124. Tastan, O., Qi, Y., Carbonell, J.G., Klein-Seetharaman, J.: Prediction of interactions between HIV-1 and human proteins by information integration. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, p. 516. NIH Public Access (2009)
125. Tato, C., Hunter, C.: Host-pathogen interactions: subversion and utilization of the NF-κB pathway during infection. Infection and Immunity **70**(7), 3311–3317 (2002)
126. Taylor, P.D., Day, T., Nagy, D., Wild, G., Andre, J.B., Gardner, A.: The evolutionary consequences of plasticity in host–pathogen interactions. Theoretical Population Biology **69**(3), 323–331 (2006)
127. Tekir, S.D., Çakır, T., Ardıç, E., Sayılırbaş, A.S., Konuk, G., Konuk, M., Sarıyer, H., Uğurlu, A., Karadeniz, İ., Özgür, A., et al.: PHISTO: pathogen–host interaction search tool. Bioinformatics **29**(10), 1357–1358 (2013)
128. 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)
129. Tobin, D.M., May, R.C., Wheeler, R.T.: Zebrafish: a see-through host and a fluorescent toolbox to probe host-pathogen interaction. PLoS Pathogens **8**(1) (2012)
130. Tomar, N., De, R.K.: A Model of an Integrated Immune System Pathway in Homo sapiens and Its Interaction with Super-antigen Producing Expression Regulatory Pathway in Staphylococcus aureus: Comparing Behavior of Pathogen Perturbed and Unperturbed Pathway. PloS One **8**(12), e80,918 (2013)
131. Torrelles, J.B., Schlesinger, L.S.: Diversity in Mycobacterium tuberculosis mannosylated cell wall determinants impacts adaptation to the host. Tuberculosis **90**(2), 84–93 (2010)
132. Vergne, I., Singh, S., Roberts, E., Kyei, G., Master, S., Harris, J., Haro, S.d., Naylor, J., Davis, A., Delgado, M., et al.: Autophagy in immune defense against Mycobacterium tuberculosis. Autophagy **2**(3), 175–178 (2006)
133. Via, A., Uyar, B., Brun, C., Zanzoni, A.: How pathogens use linear motifs to perturb host cell networks. Trends in Biochemical Sciences **40**(1), 36–48 (2015)
134. Viboud, G.I., Bliska, J.B.: Yersinia outer proteins: role in modulation of host cell signaling responses and pathogenesis. Annual Review Microbiology **59**, 69–89 (2005)
135. Vodovar, N., Acosta, C., Lemaitre, B., Boccard, F.: Drosophila: a polyvalent model to decipher host–pathogen interactions. Trends in Microbiology **12**(5), 235–242 (2004)
136. Wang, Y., Zhang, Q., Sun, M.a., Guo, D.: High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. Bioinformatics **27**(6), 777–784 (2011)
137. Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., et al.: PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic acids research p. gkt1099 (2013)
138. Weekes, M.P., Tomasec, P., Huttlin, E.L., Fielding, C.A., Nusinow, D., Stanton, R.J., Wang, E.C., Aicheler, R., Murrell, I., Wilkinson, G.W., et al.: Quantitative temporal viromics: an approach to investigate host-pathogen interaction. Cell **157**(6), 1460–1472 (2014)
139. Welch, R., Dellinger, E., Minshew, B., Falkow, S.: Haemolysin contributes to virulence of extra-intestinal E. coli infections. Nature **294**(5842), 665–667 (1981)
140. Winnenburg, R., Baldwin, T.K., Urban, M., Rawlings, C., Köhler, J., Hammond-Kosack, K.E.: PHI-base: a new database for pathogen host interactions. Nucleic Acids Research **34**(suppl 1), D459–D464 (2006)
141. Winnenburg, R., Urban, M., Beacham, A., Baldwin, T.K., Holland, S., Lindeberg, M., Hansen, H., Rawlings, C., Hammond-Kosack, K.E., Köhler, J.: PHI-base update: additions to the pathogen–host interaction database. Nucleic Acids Research **36**(suppl 1), D572–D576 (2008)
142. Wooldridge, K.: Bacterial secreted proteins: secretory mechanisms and role in pathogenesis (2009)
143. Wuchty, S.: Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens. PLoS One **6**(11), e26,960 (2011)
144. Xiang, Z., Tian, Y., He, Y., et al.: PHIDIAS: a pathogen-host interaction data integration and analysis system. Genome Biology **8**(7), R150 (2007)
145. Yang, J., Chen, L., Sun, L., Yu, J., Jin, Q.: VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Research **36**(suppl 1), D539–D542 (2008)
146. Zelle, M.R.: Genetic constitutions of host and pathogen in mouse typhoid. Journal of Infectious Diseases **71**(2), 131–152 (1942)
147. Zhou, C., Smith, J., Lam, M., Zemla, A., Dyer, M.D., Slezak, T.: MvirDB-a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Research **35**(suppl 1), D391–D394 (2007)
148. Zhou, H., Xu, M., Huang, Q., Gates, A.T., Zhang, X.D., Castle, J.C., Stec, E., Ferrer, M., Strulovici, B., Hazuda, D.J., et al.: Genome-scale RNAi screen for host factors required for HIV replication. Cell Host & Microbe **4**(5), 495–504 (2008)
149. Zychlinsky, A., Sansonetti, P.J.: Apoptosis as a proinflammatory event: what can we learn from bacteria-induced cell death? Trends in Microbiology **5**(5), 201–204 (1997)

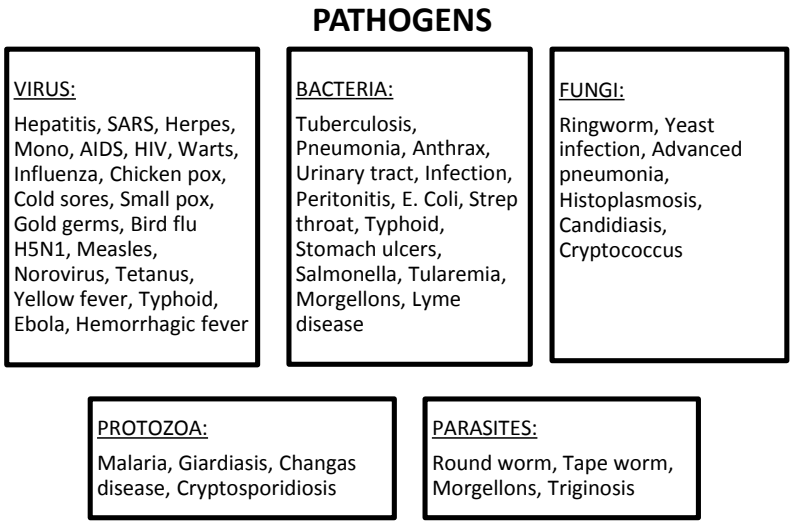

Fig. 1: Classification of some common pathogens and the list of diseases caused by them

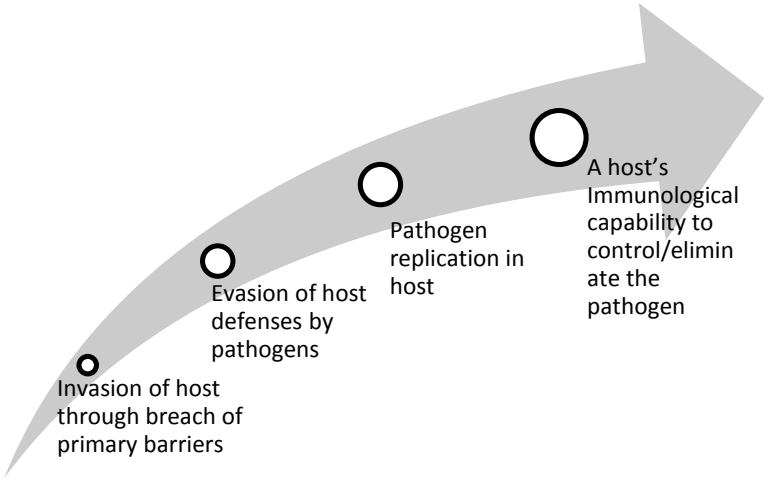

Fig. 2: Classification of Host-Pathogen Interactions

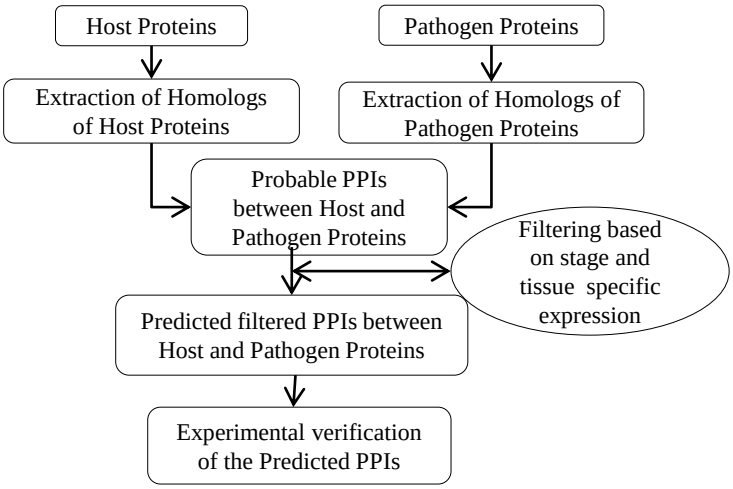

Fig. 3: Homology based predictions of host-pathogen interactions

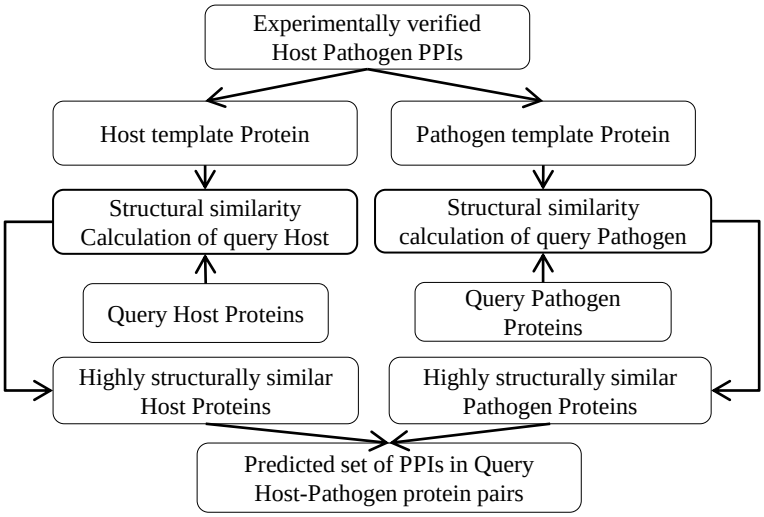

Fig. 4: Structure based predictions of host-pathogen interactions

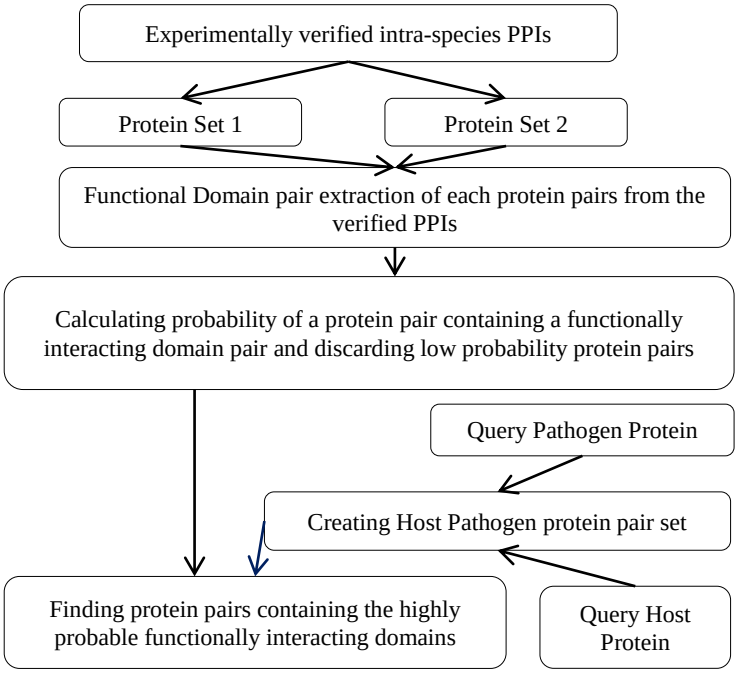

Fig. 5: Domain/motif based prediction of host-pathogen interactions

Table 1: Summary of the machine learning based tools used in the domain of host-pathogen interactions.

Machine Learning Method	Species	Reference	Domain
Random Forest Classifier	*HIV1-Homo sapiens*	Tastan *et al.* [124]	PPI
Naive Bayes Classifier	Phylum *Chlamydiae* and genera *Escheria*, *Yersinia* and *Pseudomonas*	Arnold *et al.* [4]	T3SS
Support Vector Machine	*Ralstonia solanacearum*	Wang *et al.* [136]	T3SS
Multi-task Classifier using Support Vector Machine	*Yersinia pestis*, *Francisella tularensis*, *Salmonella* and *Bacillus anthracis*	Kshirsagar *et al.* [66]	PPI
Semi Supervised Learning using Multi-layer Perceptron	*HIV1-Homo sapiens*	Yanjun *et al.* [102]	PPI
Random Forest Classifier	*Homo sapiens-Plasmodium falciparum*	Wuchty [143]	PPI
Group lasso with *l1/l2* regularization	*Homo sapiens-Salmonella*, *Homo sapiens-Yersinia*	Kshirsagar *et al.* [65]	PPI
Support Vector Machine	None	Thieu *et al.* [128]	Data Mining

Table 2: List of online repositories storing data related to host-pathogen interactions

No.	Name	URL
1	PATRIC [137]	http://patricbrc.org/portal/portal/patric/Home
2	PIG [35]	http://patricbrc.org/portal/portal/patric/HPITool
3	VirHostNet [95]	http://virhostnet.prabi.fr/
5	HPIDB [69]	http://agbase.msstate.edu/hpi/main.html
6	GPS-Prot [44]	http://gpsprot.org/
7	VirusMint [18]	http://mint.bio.uniroma2.it/virusmint/Welcome.do
8	PHIDIAS [144]	http://www.phidias.us/introduction.php
9	MvirDB [147]	http://mvirdb.llnl.gov/
10	PHI-base [140, 141]	http://www.phi-base.org/
11	PID [113]	http://pid.nci.nih.gov/
12	BioHealthBase [121]	http://www.biohealthbase.org/
13	VPDB [116]	http://www.vpdb.bicpu.edu.in/
14	VectorBase [71]	https://www.vectorbase.org/
15	EuPathDB [6]	http://eupathdb.org/eupathdb/
16	PHISTO [127]	http://www.phisto.org/
17	ViPR [99]	http://www.viprbrc.org/brc/home.spg?decorator=vipr
18	EDWIP [97]	http://cricket.inhs.uiuc.edu/edwipweb/edwipabout.htm
19	HoPaCI-db [9]	http://mips.helmholtz-muenchen.de/HoPaCI
20	VFDB [21]	http://www.mgc.ac.cn/VFs/main.htm
21	AquaPathogen X [41]	http://pubs.usgs.gov/fs/2012/3015/

Table 3: Summary of host protection and pathogen attacking mechanisms.

Host Protection Mechanism	Pathogen Attacking Mechanism
Protein-Protein Interactions (GBP galactose-binding protein)	Protein - Protein Interactions (target hub protein)
shRNAs (pathogen gene knock down)	microRNAs (protection against cellular micro-viral response,gene silencing)
Autophagy	MAM (multivariate adhesion molecule, high binding affinity with host during infection)
siRNAs (inhibit HIV-1 replication)	Pneumolysin (virulence factor)
Macrophages	Inhibition of macrophage
Restricting supply of Iron	Glial cells of host (convert it into progenitor cells then survive in the host)
None	Motif mimicry (utilized by pathogens to rewire host pathways by co-opting SLiM mediated protein interactions)
None	Biofilm formation
None	Hijacking (implant own sequence in normal sequence of host)