ImitateDB: A Resource for Domain and Motif Mimicry in Host and Pathogen Proteins

Sonali Tayal[1], Venugopal Bhatia[2], Tanya Mehrotra[1], and Sonika Bhatnagar[1]

[1]Netaji Subhas University of Technology
[2]Netaji Subhas Institute of Technology

July 22, 2021

Abstract

The host pathogen interactome can be visualized as a vast continuous network in which molecular mimicry of host proteins by pathogens constitutes a strategy to hijack the host pathways. Despite extensive work in this field, there is no dedicated resource for mimicked domains and motifs in host pathogen interactome. In this work, we collated all the data regarding the experimental host pathogen (HP) and host-host (HH) protein-protein interactions (PPIs). The domains and sequence linear motifs of the proteins were annotated using CD Search and ScanProsite. Host and pathogen proteins with a shared host interactor and similar domain/motif constitute a linear pair. A linear pair that exhibits global structural domain similarity (Domain linear pair; DLP) or local sequence motif similarity (Motif Linear Pair; MLP) has a high probability of being co-expressed and co-localized. 2,06,449 DLPs and 38,45,643 MLPs were identified in 50,812 experimental HP-PPIs and organized in a web- based resource, ImitateDB, accessible at http://imitatedb.sblab-nsit.net. ImitateDB provides user-friendly access to the mimicry data. It can be queried by protein UniProt ID, pathogen, domain, motif, or interaction detection method. The results are externally integrated using hyperlinked domain PSSM ID, motif ID and protein ID. Kinase, UL36, Smc and DEXDc were frequent DLP domains whereas Protein Kinase C, Casein Kinase 2, glycosylation and myristoylation sites were frequent MLP motifs. Novel DLP domains SANT, Tudor, PhoX and MLP motifs Microbodies C-terminal targeting signal, Ubiquitin-interacting motif and Lipocalin signature were proposed. ImitateDB constitutes a resource for researchers in the field of infectious diseases and microbiology.

Introduction

In nature, Host-pathogen Protein-Protein interactions (HP-PPIs) are highly complex, ubiquitous and fairly essential for elucidation of infectious diseases (1). During this interaction, there is a continuous cross talk between pathogens and their hosts that is mediated by a variety of effectors including proteins, small molecules, metabolites, and regulatory RNAs(2, 3). Pathogenesis involves interactions between the signalling networks of the host and pathogen. Recent studies regarding HP-PPIs focus on the mechanisms employed by pathogens to hijack and exploit the host immune system for their own survival. Processes for molecular mimicry have evolved to enable the proteins of pathogens to imitate the host proteins in order to disrupt their interactions and disturb the signalling pathways (4). Thus, the interacting pathways and proteins of the pathogen may be conceived to be in a continuum with those of the host.

Mimicry of host antigenic determinants as a survival mechanism was described early in parasites (5). A pathogen's ability to mimic the host components may be achieved by two distinct mechanisms. The first one is where the host genes are acquired by the pathogen through horizontal gene transfer. An example of this is the acquisition of complement escape regulators by pathogenic bacteria like *Echinococcus granulosus* (6) and *Onchocerca volvulus* (7). The second mechanism is where both host and pathogen genes evolved independently and ended up having similar structures with different function i.e. underwent convergent evolution(8). A well-known example of this is the *Yersinia pseudotuberculosis* effector protein, invasin, that

structurally mimics the integrin-binding surface of the protein fibronectin (9). While Horizontal gene transfer leads to a detectable homology between the pathogen and host proteins (10, 11), convergent evolution is likely to modulate local similarity between the proteins of pathogen and host as depicted by sharing of motifs (12). The local similarities between epitopes from the pathogen/infectious agents and antigens present in the host can also lead to autoimmune diseases (13-17).

Molecular mimicry can operate at four distinct levels; (i) Similarity in both sequence and structure of a full-length protein or a functional domain as displayed by molecular mimicry between *Legionella pneumophila* , *Chlamydia trachomatis* and *Burkholderia thailandensis* SET-domain containing proteins with host proteins (18), (ii) only the structural similarity without an apparent sequence similarity as detected in case of several bacterial and viral pathogens that eventually evolved to structurally mimic host ligands, though the sequence similarity between pathogen molecules and the mimicked host ligands was low (19), (iii) similarity in the sequence of a short linear motif. An example of motif mimicry is displayed by the WxxxE motif in many bacterial Guanine Exchange Factors, such as EspM2 and Map in E. coli and also SifA of Salmonella (20, 21). Motifs have the ability to tolerate mutations and can evolve rapidly to alter interactions with the host (22), (iv) Similarity of only the binding site architectures (interface mimicry) without sequence homology is displayed by human fibronectin and *Y. pseudotuberculosis* invasin binding to human integrin (9, 11). These proteins display similarity in the chemical properties at the binding site in the absence of sequence and structural homology.

The existing methods of detection of mimicry are simply based on identifying sequence or structure similarity. A previously available database, namely mimicDB (8) provides information about molecular mimicry proteins or epitopes involved in a limited number of human parasites. Another database miPepBase (23) lists the experimentally verified mimicry peptides involved in auto-immune disease. However, a wide range of domains and motifs are recruited by pathogens to mimic and hijack the host cell machinery for its survival (20, 21, 24-26). A computational pipeline using pBLAST against the human proteome has also been implemented for the prediction of the molecular mimicry candidates in bacterial pathogens (27). However, sequence-based methods for discovery of protein mimics may not be adequate as they are dependent on the level of recognizable homology between the host and pathogen proteins. Structure-based methods are more suitable for recognizing remote similarity while motif-based methods are suitable for recognizing localized regions of similarity between proteins. Pathogenic bacteria are likely to target the host proteins by imperfectly mimicking the host interface (28). An interface mimicry-based method, the HMI-PRED server (29) carries out structural prediction of given HP-PPIs. However, it is limited due to the requirement of the structure of the microbial protein involved in mimicry.

Similarity between motifs and domains of the host and pathogen proteins does not necessarily indicate their actual interaction. This is further dependent on the proteins having simultaneous expression and being present in the same cellular compartment. However, analysis of the PPIs in yeast and human showed that a large majority of the interactions occur between proteins in the same subcellular compartment (30, 31). Studies have also shown that functionally related or interacting proteins from the same pathways share Gene Ontology, and also usually constitute a higher co-expression score (32, 33). Also, imitation of host proteins by the pathogen essentially works by imitation and competing with endogenous (host–host) interactions(34, 35). We therefore hypothesize that resemblance between the experimentally validated host and pathogen interactors of the same host protein increases the confidence in the identification of molecular mimicry candidates due to colocalization and co-expression of the interacting protein pairs. This is shown schematically in Figure 1a for global structural similarity (domain linear pair or DLP) and Figure 1b for local sequence similarity (motif linear pair or MLP). Delineating the DLPs and MLPs also provides information about the host interactions that are likely to be disrupted by pathogen protein mimicry.

In this work, we collated the entire set of experimental HP-PPIs from interaction databases in order to compute their DLPs and MLPs, which were organized in the form of a publicly available database, ImitateDB available online at http://imitatedb.sblab-nsit.net. The ImitateDB resource can help researchers to search for organism-wise mimicry patterns prominent in the host pathogen interactome. It houses 2,06,449 DLPs and

38,45,643 MLPs. Out of the total 61,215 HP-PPIs collated, 1,549 and 49,266 were found to be characterized by imitated domains and motifs. Several novel potential domain mimics include SANT (Swi3, Ada2, N-Cor, and TFIIIB) DNA binding domain, Tudor and PhoX homology domain while some of the novel motif mimics identified are Microbodies C-terminal targeting signal, Ubiquitin-interacting motif and Lipocalin signature. Specific domains or motifs imitated commonly by a large number of pathogens are likely to be responsible for microbial virulence suitable for drug/vaccine targeting. Thus, ImitateDB constitutes a source of information for molecular imitation in HP-PPIs for researchers in the field of infectious diseases and microbiology.

Methodology

HP-PPI Data collection and cleaning

The information regarding the host HP-PPIs for the database was collected from different HP-PPIs databases namely BioGrid(36), PHISTO(37), HPIDb(38), MINT(39), IntAct(40), MPIDB(41), UniProt(42), VirHostNet(43), MatrixDB(44), I2D(45), DIP(46) and InnateDB(47). The data obtained from these sources included information about i) UniProt accession numbers, ii) Gene symbols, iii) UniProt entry names, iv) Gene symbols for the interacting proteins of pathogen and human host, v) Corresponding pathogen names for all pathogen proteins, vi) Pathogen taxon IDs, and vii) Experimental method of interaction detection for each unique interaction.

The UniProt accession number was used as a unique identifier for the proteins extracted from different sources to maintain uniformity in the data. The pathogen names from different databases were also examined for variations in syntax/nomenclature and were converted into a uniform name using UniProt Taxon identifier. The duplicate entries were removed from the data to avoid redundancy and the obsolete entries were either removed or converted into secondary uniport accession if available.

Host-Host Interaction (HHI) data collection

The information of the first interactor proteins of the host proteins were also collected using UniProt. This data was further processed using bioDBnet database (48) to filter out the various non-human interactors of the host proteins and retain only the human interactors of the host proteins.

Domain annotation

Each interacting pathogen protein, the corresponding host protein and host first interactor were examined for structural domain or local sequence motif similarity. Domain annotation was carried out using the NCBI Batch Conserved domain (CD) Search (49). CD-Search is a sensitive method that constructs the models of structurally conserved domain families based on multiple sequence alignments which are converted into position-specific scoring matrices (PSSM). It scans protein query sequences against these matrices with Reverse Position Specific-BLAST, a variant of the Psi-BLAST algorithm. Structure being more conserved than sequence, CD Search identifies remote similarity between pathogen and host proteins. The domain information was collected in the form of a unique PSSM ID and Domain short name for every unique domain family of all the proteins.

Motif annotation

The motifs were identified using the Expasy ScanProsite (50) API (Application Programming Interface) with the help of a Python script. ScanProsite tool takes into account the ProRules which focuses on extreme similarity at only a few biologically significant sites of well-chosen residues i.e motifs or signatures and also defines the position of these structurally and functionally crucial amino acids. The rules are further used internally, by the Swiss-Prot for annotation of protein query sequences. The motif information for each protein was collected in the form of ScanProsite IDs, motif name and pattern.

Identification of DLPs and MLPs

The pathogen proteins and host proteins interacting with the same protein were compared for shared structural domains (global similarity) or linear sequence motifs (local similarity) using a Python script. Python

libraries such as NumPy and pandas were used to identify the DLPs and MLPs between the pathogen and host first interactors.

Database and web interface development

After the completion of data processing and formation of domain and motif linear pairs, a database was created using MySQL to house the schema and various tables through MySQL Command line client server. Further, an MVC (Model View Controller) web application was created in Node.js using the Express framework. The front end of the web application was developed using HTML (HyperText Markup Language), EJS (Embedded JavaScript Templating - to render HTML with our own set of variables) and CSS (Cascading Style Sheets) with JavaScript and jQuery being employed for implementing the various methods and functions. Further we used Redis for queuing tasks and MongoDB for caching and storing session cookies.

Data analysis and Visualization

Analysis of the database was done using MySQL, Microsoft Excel, R studio and Tableau. MySQL was used to sort the entities and quantitatively measure the frequency of each entity in the entire data. The graphs and highlight tables depicting the analysis were rendered using Microsoft Excel and Tableau. The high degree pathogen and host proteins were identified using an R script. Pathway enrichment and gene ontology analysis of the host proteins was carried out using PANTHER (51).

Results and Discussion

ImitateDB offers comprehensive information about the molecular mimicry candidates among the interacting pathogen and host proteins in the form of DLPs and MLPs which linear provide information about the matched domains and motifs between the pathogen proteins and first interactor proteins of the respective interacting host proteins. The information in the database is divided into different categories according to different categories of pathogen like Virus, Bacteria, and Fungi. The pathogens belonging to Protozoa, Amoebozoa and Archaea are found under the "Others" category. The schematic of the database is shown in Figure 2. It depicts the basic pipeline followed for all the search options and the basic work flow for the development of the database as well as frequency of primary entities in the database.

HP-PPI data

After collection and processing of the HP-PPI data, it was found that viruses have the highest number of reported HP-PPIs among the different pathogen categories. After data collection and cleaning, 5,569 pathogen proteins from 630 organisms interacted with 10,078 host proteins with 61,215 HP-PPIs. There were 49,249 reported unique HP-PPIs of viral origin. In comparison, reported bacterial HP-PPIs were 10,080 while those from other organisms were even fewer. Further, 11,657 host first interactors having 1,03,120 interactions with the host proteins were retrieved as described in the Methodology. Domains and motifs of these first interactors were identified and compared with those present in the pathogen proteins.

Domain and Motif Annotation

A total of 68,838 and 4,78,710 domains were annotated in the pathogen and host interactor proteins respectively. A total of 31,594 motifs were annotated in the pathogen and 79,944 in the host proteins respectively. Since, each domain or motifs can occur multiple times in different proteins, the number of unique conserved domains and motif IDs were counted. There were 17,465 and 25,245 unique domains in pathogens and host interactor proteins respectively while the total number of unique motifs were 1046 and 1661 respectively.

MLPs are more numerous in comparison with DLPs

Out of the 5,569 pathogen proteins from 630 pathogens, a total of 5,255 proteins from 612 pathogens were involved in mimicry of the host interactor proteins using similar domains or motifs. The DLPs in the entire database were found to be 2,06,449 whereas the MLPs were found to 38,45,643. The number of DLPs and MLPs for each pathogen category are listed in Table 1. Viruses showed the highest number of DLPs and MLPs, likely to be due to the preponderance of virus HP-PPIs in the data.

Interestingly, of the total 61,215 HP-PPIs reported, only 1,549 were found to be characterized by domain mimicry whereas 49,266 were found to characterized by mimicked motifs. The total number of HP-PPIs, the fraction of HP-PPIs characterized by mimicked domains and by motifs were compared pathogen categories and are shown in Figure 3. Motif mimicry dominates in number over domain mimicry across the known HP-PPIs across all pathogen categories. Interestingly, as evident from Table 1, a large number of DLPs were also found in viruses (1,14,899), but were concentrated in only 822 HP-PPIs (or 1.7%) of the entire viral interactome as depicted in Figure 3. Conversely, in case of Fungi, very few HP-PPIs have been reported but still a large proportion of them were found to be characterized by mimicked candidates. Previous studies have reported the extensive use of motif mimicry by viral proteomes (25, 52-54). A reason for this monopoly can be that viruses being obligately intra cellular, incapable of synthesizing the DNA or RNA (25), and having a rapidly evolving genome need to hijack the major host processes which includes various metabolic and cellular signalling pathways. Our data indicates that molecular mimicry may be comparatively much more frequent in case of fungi. However, our results about frequency should be treated with caution as the mapping of the host pathogen interactome is far from complete.

Domain mimicry

Out of the 5,569 pathogens proteins from 630 pathogens, 607 proteins from 146 pathogens made DLPs with the host interactor proteins as indicated in the schematic Figure 1. Since there were multiple instances of every mimicked domain, we looked for unique domain types. There existed 3040 types of unique cdd domains shared by both pathogens and host. The largest number of DLPs were found for the Serine Threonine Protein Kinase US3 (UniProt ID: P04413) from *Human Herpesvirus 1 Strain 17 (HHV-1)* with 61,609 DLPs. The top 10 pathogens involved in molecular mimicry along with the number of DLPs are shown in Table 2. Two viral pathogens with the maximum number of DLPs were *HHV-1*and *Rous sarcoma virus strain Schmidt-Ruppin A* . In case of bacteria, *Legionella pneumophila subsp. pneumophila (strain Philadelphia 1 / ATCC 33152 / DSM 7513)* was found to have the largest number and widest diversity of host-like domains (Table 2). This opportunistic human bacterial pathogen has previously been reported to be highly involved in molecular mimicry of host proteins (24, 55).

The top 10 most frequently observed mimicked domains are shown in Figure 4a. PHA03247 (large tegument protein UL36) was the most frequent among DLPs. UL36 is an important domain family of tegument protein of Herpes Simplex Virus (HSV) that is crucial for virus host interaction and host immune evasion (56). UL36 is found to be colocalized with host and viral membrane proteins and aid in the assembly and cell entry of HSV(57). The top 10 most frequently occurring mimicked domains in different organism categories are shown in Table 3. A conserved domain family found to be potentially mimicked by viruses was DEAD-like helicases domain superfamily. The DEAD-box helicases bear a common D-E-A-D motif and is an emerging class of host proteins being mimicked by viruses for infections (58). Bacterial, viral and fungal conserved domains found in most frequently in DLPs were Rad50 ATPase and SbcC. Rad50 ATPase and SbcC are both involved in DNA repair pathways and are highly conserved among eukaryotes (humans and fungi), bacteria and viruses as well (59, 60).This way, the pathogens seem to have captured DNA repair proteins from their hosts to aid their own replication and survival by disrupting the host DNA repair pathways (61, 62).

Another important mimicked domain found in our data is Glycogen Synthase Kinase-3 (GSK-3) domain. Bacterial pathogen such as *Helicobacter pylori* has been found to divert the host signalling pathways such as WNT signalling by targeting the host GSK-3(61).

The predominantly occurring fungal pathogen found to mimic the largest number of host-like domains was found to be *Saccharomyces cerevisiae S288C* . In case of Others category, *Dictyostelium discoideum* is the predominant pathogen imitating the maximum number of domains. The pathogens with the highest number of DLPs and MLPs in different pathogen categories, i.e., virus, bacteria, fungi, and others are listed in Supplementary data Tables S1, S2, S3 and S4 respectively.

Motif mimicry

Out of the 5,569 pathogen proteins from 630 pathogens, 5,255 proteins from 610 pathogens made MLPs

with the host interactor proteins as indicated in the schematic Figure 1. However, only 239 unique motifs were found to be mimicked by pathogens. Since each pathogen can mimic motifs from multiple interactors, the largest number of MLPs were found for the Polymerase basic protein 2 from *Influenza A virus strain A/Wilson-Smith/1933 H1N1 (A/Wilson-Smith/33/H1N1),* with 35,385 MLPs. The average number of MLPs for a protein is 732.Amongst viral pathogens*, A/Wilson-Smith/33/H1N1* had the maximum MLPs whereas in the bacterial interactome, *Yersinia pestis* had the maximum MLPs. The Top 10 pathogens by the count of MLPs are listed in Table 4.

Table 2 and 4 showed that *S. cerevisiae S288c* had the maximum count of DLPs and MLPs even while the total number of reported HP-PPIs were very low in comparison with virus or bacteria. This can be attributed to the fact that yeasts, being eukaryotes are quite similar to humans in terms of genes and other cellular pathways. It has been observed that the genes that regulate cellular processes in humans have equivalents that control cell division in yeasts as well which makes it very easy for pathogenic yeast species to alter the host cellular machinery (63). Therefore, this study has unravelled the potential mimicry candidates in fungal pathogens which was not well established till now.

The total count for the top 10 most frequently occurring motifs in the database is shown in Figure 4b. The predominance of phosphorylation sites for Protein kinase C (PKC) phosphorylation site and casein kinase II (CK2) phosphorylation site can be observed from the figure. PKC and CK2 family of serine/threonine kinases plays essential roles in hijacking multiple signalling pathways in humans leading to many viral infections (64). Tyrosine phosphorylation has been proved to be an important process for pathogenesis as well as immune responses after the underlying revelation of a bacterial tyrosine phosphatase (65). There have been instances where both extracellular as well as intracellular bacteria secreted several proteins that mimicked the function of their analogous eukaryotic like proteins and hijacked the tyrosine phosphorylation pathway (66). Additionally, sites for N-myristoylation, Amidation site, and N-glycosylation could be seen in all the organism categories. Several instances have showed the contribution of post translational modification (PTM) sites in microbial infection and cellular processes (67, 68).

The top 10 most frequent motifs in every pathogen category are listed in Table 5. N- glycosylation was a frequently occurring motif known to be an important modification used by several pathogen proteins (specifically viral glycoproteins) to evade the human immune system (69, 70). The envelope proteins of viruses like HIV-1 are heavily glycosylated and can provide camouflage against the human proteins, leading to alteration of immune recognition (71, 72). Protein N-myristoylation site is another conserved PTM of proteins involved in a variety of different physiological processes like cell proliferation and differentiation, cell survival, and cell death(73). Also, several myristoylated proteins have been found to have prominent roles in cellular signalling pathways (74) and the myristoylation motif has been found to be mimicked by viral and bacterial proteins (25, 75).

Additionally, several other commonly mimicked motifs in our data were ABC transporters family signature motif, Q motif, ATP/GTP-binding site motif A (P-loop), arginine-rich motif, ubiquitination site and prenyl group binding site. The ABC transporters family signature motif is a conserved sequence (LSGGQ) present in the Nucleotide binding domain (NBD) of all ABC transporters and is primarily required for substrate transport (76). The pathogens can mimic this motif to disturb the transportation pathways of the host. Q motif is a part of conserved helicases (involved in DNA dynamics) (77) and might help the pathogens to hijack the host machinery associated with DNA replication, recombination, transcription, and repair. The highlight table depicting the number of MLPs characterized by top 20 mimicked motifs for the top 20 pathogens is shown as Supplementary Table S5

Linear pairs in highly interacting pathogen proteins and host proteins

Several previous studies have shown that essentiality and pathogen fitness are correlated with high number of interactions (78, 79). Therefore, the number of DLPs and MLPs in the top 10 highly interacting pathogen proteins and host proteins were examined and are listed in Table 6 and Table 7 respectively. The top 10 highly interacting pathogen proteins were of viral origin and predominantly formed MLPs, indicating that

local sequence similarities predominate the mimicry of host proteins by pathogens. Among host proteins, a few had a very high number of DLPs. As shown in Table 7, Nuclear factor NF-kappa-B p105 subunit was a part of 491 while Cellular tumor antigen p53 was a part of 180 DLPs.

Chemokine & cytokine, Wnt, CCKR, EGR receptor and PDGF signalling pathways are enriched in host proteins of linear pairs

The enriched pathways and processes of the host primarily imitated by pathogens during infection were annotated. Apart from specific pathways for some autoimmune diseases such as Huntington and Alzheimer, Chemokine & cytokine, Wnt, CCKR (Cholecystokinin receptor), EGF (Epidermal Growth Factor) receptor, PDGF (Platelet-derived growth factor) signalling pathways and T-cell and B-cell activation pathways were enriched among the host proteins constituting the DLPs and MLPs (Figure 5). Thus, signalling pathways of the hosts are most commonly targeted by the pathogens for molecular mimicry.

The host proteins were further annotated using gene ontologies. The enriched cellular compartments found were cellular anatomical entities, protein-containing complex and intracellular. As shown in Supplementary Figure S1 and S2, the enriched molecular functions found were binding, catalytic activity, molecular function regulator while enriched biological processes found were cellular process, metabolic process and signalling.

Selected novel domain and motif mimicry candidates

Several novel candidate mimicry domains like SANT, TCP-1 and Tudor in pathogens were identified from analysis of the ImitateDB data. Some of the novel domain mimics identified in different pathogens along with their functions are shown in Table 8. Microbodies C-terminal targeting signal, RNA recognition motif and Ubiquitin-interacting motif were novel motif mimic candidates. Selected of the novel motif mimics identified in different pathogens along with their functions are shown in Table 9.

The ImitateDB web interface

The web interface for the ImitateDB database provides a user-friendly access to the data and allows the user to search for information about DLPs and MLPs using multiple search options. The web interface has a home page delineating the basic concept a radio button to choose between DLPs and MLPs. After this selection, the interface provides the next radio button to choose among the different categories of pathogens, namely: Virus, Bacteria, Fungi, Others. In each category, the database can be searched by Organism, Pathogen Protein ID, Host Protein ID, Interaction Detection method, Host interactor protein ID, Matched Domain PSSM ID, Domain Short name, Matched Motif ID, Motif Name or pattern. For easier searching, selection of the category and subcategory leads to the population of a drop-down menu with available options. Additionally, the user can enter a keyword to retrieve the required data. After this selection, the user needs to enter the captcha correctly to fetch the results.

An expanded view of the search panel in the database is shown in Figure 6a. The results are displayed in the form of a table that can be downloaded. The download feature for bulk files has been restricted due to download constraints. Queries yielding results between 10,000 - 1,00,000 records, the user is provided with the results by email using an in-built mailer (as shown in Figure 6b), that pops up after clicking the download button.

Conclusion

Pathogens have evolved a large number of ways of impersonating human proteins over a period of time. Pathogen proteins may mimic domain and motifs of the first interactors of the interacting host protein, and thus disturb the various signalling pathways of the humans. Imitate DB lists all the mimicked domains and motifs among HP-PPIs. The integration of HP-PPI data with domain and domain/motif mimicry is likely to predict the mimicry candidates with higher confidence. An exception to this would be when the proteins in a DLP or MLP have distinct temporal or spatial interaction. The limitation of our method is that the mimicry candidates will only be identified for those organisms and proteins for which the experimental PPIs have been reported in the databases. The curated information in ImitateDB will help in identifying frequent, unique,

and novel mimicry domains and motifs among the interacting hosts and pathogens. Additionally, MLPs or DLPs allows us to easily identify and model the host protein motif or domain at which the competition for binding sites is taking place. The disruption of these HP-PPIs can be regarded as a strategy for developing novel broad-spectrum therapeutics against multiple infectious diseases.

Author Contributions

ST carried out data cleaning, enrichment and organisation, development of the database, analysis, and manuscript preparation. VB developed the backend and front end of the web interface. TM carried out data acquisition and cleaning. S.B. was involved in conception, design, analysis and supervision of the study. The manuscript was reviewed by all the authors.

Data Availability

The authors confirm that the data supporting the findings of this study is available at http://imitatedb.sblab-nsit.net.

Conflict of interest

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Dean Southwood SR. Host-Pathogen Interactions. Encyclopedia of Bioinformatics and Computational Biology. 2019;3:103-12.

2. Olive AJ, Sassetti CM. Metabolic crosstalk between host and pathogen: sensing, adapting and competing. Nat Rev Microbiol. 2016;14(4):221-34.

3. Moreira D, Estaquier J, Cordeiro-da-Silva A, Silvestre R. Metabolic Crosstalk Between Host and Parasitic Pathogens. Exp Suppl. 2018;109:421-58.

4. Guven-Maiorov E, Tsai CJ, Nussinov R. Pathogen mimicry of host protein-protein interfaces modulates immunity. Semin Cell Dev Biol. 2016;58:136-45.

5. RT D. Molecular mimicry: antigen sharing by parasite and host and its consequences. Am Nat 1964;98.

6. Diaz A, Ferreira A, Sim RB. Complement evasion by Echinococcus granulosus: sequestration of host factor H in the hydatid cyst wall. J Immunol. 1997;158(8):3779-86.

7. Meri T, Jokiranta TS, Hellwage J, Bialonski A, Zipfel PF, Meri S. Onchocerca volvulus microfilariae avoid complement attack by direct binding of factor H. J Infect Dis. 2002;185(12):1786-93.

8. Ludin P, Nilsson D, Maser P. Genome-wide identification of molecular mimicry candidates in parasites. PLoS One. 2011;6(3):e17546.

9. Hamburger ZA, Brown MS, Isberg RR, Bjorkman PJ. Crystal structure of invasin: a bacterial integrin-binding protein. Science. 1999;286(5438):291-5.

10. Sallee NA, Rivera GM, Dueber JE, Vasilescu D, Mullins RD, Mayer BJ, et al. The pathogen protein EspF(U) hijacks actin polymerization using mimicry and multivalency. Nature. 2008;454(7207):1005-8.

11. Stebbins CE, Galan JE. Structural mimicry in bacterial virulence. Nature. 2001;412(6848):701-5.

12. Dong Yu ZY, Yuan Jin, Jing Zhou, Hongguang Ren, Mingda Hu, Beiping Li, Wei Zhou, Long Liang, Junjie Yue. Evolution of bopA Gene in Burkholderia: A Case of Convergent Evolution as a Mechanism for Bacterial Autophagy Evasion. BioMed Research International. 2016;2016:7.

13. Zabriskie JB, Freimer EH. An immunological relationship between the group. A streptococcus and mammalian muscle. J Exp Med. 1966;124(4):661-78.

14. Cusick MF, Libbey JE, Fujinami RS. Molecular mimicry as a mechanism of autoimmune disease. Clin Rev Allergy Immunol. 2012;42(1):102-11.

15. Venigalla SSK, Premakumar S, Janakiraman V. A possible role for autoimmunity through molecular mimicry in alphavirus mediated arthritis. Sci Rep. 2020;10(1):938.

16. McClain MT, Heinlen LD, Dennis GJ, Roebuck J, Harley JB, James JA. Early events in lupus humoral autoimmunity suggest initiation through molecular mimicry. Nat Med. 2005;11(1):85-9.

17. Gross DM, Forsthuber T, Tary-Lehmann M, Etling C, Ito K, Nagy ZA, et al. Identification of LFA-1 as a candidate autoantigen in treatment-resistant Lyme arthritis. Science. 1998;281(5377):703-6.

18. Escoll P, Mondino S, Rolando M, Buchrieser C. Targeting of host organelles by pathogenic bacteria: a sophisticated subversion strategy. Nat Rev Microbiol. 2016;14(1):5-19.

19. Drayman N, Glick Y, Ben-nun-shaul O, Zer H, Zlotnick A, Gerber D, et al. Pathogens use structural mimicry of native host ligands as a mechanism for host receptor engagement. Cell Host Microbe. 2013;14(1):63-73.

20. Bulgin R, Raymond B, Garnett JA, Frankel G, Crepin VF, Berger CN, et al. Bacterial guanine nucleotide exchange factors SopE-like and WxxxE effectors. Infect Immun. 2010;78(4):1417-25.

21. Huang Z, Sutton SE, Wallenfang AJ, Orchard RC, Wu X, Feng Y, et al. Structural insights into host GTPase isoform selection by a family of bacterial GEF mimics. Nat Struct Mol Biol. 2009;16(8):853-60.

22. Hraber P, O'Maille PE, Silberfarb A, Davis-Anderson K, Generous N, McMahon BH, et al. Resources to Discover and Use Short Linear Motifs in Viral Proteins. Trends Biotechnol. 2020;38(1):113-27.

23. Garg A, Kumari B, Kumar R, Kumar M. miPepBase: A Database of Experimentally Verified Peptides Involved in Molecular Mimicry. Front Microbiol. 2017;8:2053.

24. Mondino S, Schmidt S, Buchrieser C. Molecular Mimicry: a Paradigm of Host-Microbe Coevolution Illustrated by Legionella. mBio. 2020;11(5).

25. Davey NE, Trave G, Gibson TJ. How viruses hijack cell regulation. Trends Biochem Sci. 2011;36(3):159-69.

26. Minton K. Plant immunity: Host mimicry of pathogen virulence targets. Nat Rev Immunol. 2015;15(7):401.

27. Doxey AC, McConkey BJ. Prediction of molecular mimicry candidates in human pathogenic bacteria. Virulence. 2013;4(6):453-66.

28. de Groot NS, Torrent Burgas M. Bacteria use structural imperfect mimicry to hijack the host interactome. PLoS Comput Biol. 2020;16(12):e1008395.

29. Guven-Maiorov E, Hakouz A, Valjevac S, Keskin O, Tsai CJ, Gursoy A, et al. HMI-PRED: A Web Server for Structural Prediction of Host-Microbe Interactions Based on Interface Mimicry. J Mol Biol. 2020;432(11):3395-403.

30. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol. 2000;18(12):1257-61.

31. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet. 2006;38(3):285-93.

32. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics. 2005;6:227.

33. Durmus Tekir S, Cakir T, Ulgen KO. Infection Strategies of Bacterial and Viral Pathogens through Pathogen-Human Protein-Protein Interactions. Front Microbiol. 2012;3:46.

34. Franzosa EA, Xia Y. Structural principles within the human-virus protein-protein interaction network. Proc Natl Acad Sci U S A. 2011;108(26):10538-43.

35. Yapici-Eser H, Koroglu YE, Oztop-Cakmak O, Keskin O, Gursoy A, Gursoy-Ozdemir Y. Neuropsychiatric Symptoms of COVID-19 Explained by SARS-CoV-2 Proteins' Mimicry of Human Protein Interactions. Front Hum Neurosci. 2021;15:656313.

36. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(Database issue):D535-9.

37. Durmus Tekir S, Cakir T, Ardic E, Sayilirbas AS, Konuk G, Konuk M, et al. PHISTO: pathogen-host interaction search tool. Bioinformatics. 2013;29(10):1357-8.

38. Ammari MG, Gresham CR, McCarthy FM, Nanduri B. HPIDB 2.0: a curated database for host-pathogen interactions. Database (Oxford). 2016;2016.

39. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al. MINT: the Molecular INTeraction database. Nucleic Acids Res. 2007;35(Database issue):D572-4.

40. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, et al. IntAct: an open source molecular interaction database. Nucleic Acids Res. 2004;32(Database issue):D452-5.

41. Goll J, Rajagopala SV, Shiau SC, Wu H, Lamb BT, Uetz P. MPIDB: the microbial protein interaction database. Bioinformatics. 2008;24(15):1743-4.

42. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158-D69.

43. Navratil V, de Chassey B, Meyniel L, Delmotte S, Gautier C, Andre P, et al. VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. Nucleic Acids Res. 2009;37(Database issue):D661-8.

44. Launay G, Salza R, Multedo D, Thierry-Mieg N, Ricard-Blum S. MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. Nucleic Acids Res. 2015;43(Database issue):D321-7.

45. Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol. 2007;8(5):R95.

46. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. Nucleic Acids Res. 2000;28(1):289-91.

47. Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, et al. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Mol Syst Biol. 2008;4:218.

48. Mudunuri U, Che A, Yi M, Stephens RM. bioDBnet: the biological database network. Bioinformatics. 2009;25(4):555-6.

49. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 2011;39(Database issue):D225-9.

50. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues

in proteins. Nucleic Acids Res. 2006;34(Web Server issue):W362-5.

51. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, et al. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 2005;33(Database issue):D284-8.

52. Duro N, Miskei M, Fuxreiter M. Fuzziness endows viral motif-mimicry. Mol Biosyst. 2015;11(10):2821-9.

53. Via A, Uyar B, Brun C, Zanzoni A. How pathogens use linear motifs to perturb host cell networks. Trends Biochem Sci. 2015;40(1):36-48.

54. Garamszegi S, Franzosa EA, Xia Y. Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks. PLoS Pathog. 2013;9(12):e1003778.

55. Gomez-Valero L, Rusniok C, Carson D, Mondino S, Perez-Cobas AE, Rolando M, et al. More than 18,000 effectors in the Legionella genus genome provide multiple, independent combinations for replication in human cells. Proc Natl Acad Sci U S A. 2019;116(6):2265-73.

56. Newcomb WW, Brown JC. Structure and capsid association of the herpesvirus large tegument protein UL36. J Virol. 2010;84(18):9408-14.

57. Schipke J, Pohlmann A, Diestel R, Binz A, Rudolph K, Nagel CH, et al. The C terminus of the large tegument protein pUL36 contains multiple capsid binding sites that function differently during assembly and cell entry of herpes simplex virus. J Virol. 2012;86(7):3682-700.

58. Meier-Stephenson V, Mrozowich T, Pham M, Patel TR. DEAD-box helicases: the Yin and Yang roles in viral infections. Biotechnol Genet Eng Rev. 2018;34(1):3-32.

59. G A Cromie JCC, D R Leach. Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans. Mol Cell. 2001;8:1163-74.

60. Yoshida T, Claverie JM, Ogata H. Mimivirus reveals Mre11/Rad50 fusion proteins with a sporadic distribution in eukaryotes, bacteria, viruses and plasmids. Virol J. 2011;8:427.

61. Gagnaire A, Nadel B, Raoult D, Neefjes J, Gorvel JP. Collateral damage: insights into bacterial mechanisms that predispose host cells to cancer. Nat Rev Microbiol. 2017;15(2):109-28.

62. Lilley CE, Schwartz RA, Weitzman MD. Using or abusing: viruses and the cellular DNA damage response. Trends Microbiol. 2007;15(3):119-26.

63. Cazzanelli G, Pereira F, Alves S, Francisco R, Azevedo L, Dias Carvalho P, et al. The Yeast Saccharomyces cerevisiae as a Model for Understanding RAS Proteins and their Role in Human Tumorigenesis. Cells. 2018;7(2).

64. Keating JA, Striker R. Phosphorylation events during viral infections provide potential therapeutic targets. Rev Med Virol. 2012;22(3):166-81.

65. Salomon D, Orth K. What pathogens have taught us about posttranslational modifications. Cell Host Microbe. 2013;14(3):269-79.

66. Selbach M, Paul FE, Brandt S, Guye P, Daumke O, Backert S, et al. Host cell interactome of tyrosine-phosphorylated bacterial proteins. Cell Host Microbe. 2009;5(4):397-403.

67. Kumar R, Mehta D, Mishra N, Nayak D, Sunil S. Role of Host-Mediated Post-Translational Modifications (PTMs) in RNA Virus Pathogenesis. Int J Mol Sci. 2020;22(1).

68. Wimmer P, Schreiner S. Viral Mimicry to Usurp Ubiquitin and SUMO Host Pathways. Viruses. 2015;7(9):4854-72.

69. Crispin M, Doores KJ. Targeting host-derived glycans on enveloped viruses for antibody-based vaccine design. Curr Opin Virol. 2015;11:63-9.

70. Crispin M, Ward AB, Wilson IA. Structure and Immune Recognition of the HIV Glycan Shield. Annu Rev Biophys. 2018;47:499-523.

71. Wagh K, Kreider EF, Li Y, Barbian HJ, Learn GH, Giorgi E, et al. Completeness of HIV-1 Envelope Glycan Shield at Transmission Determines Neutralization Breadth. Cell Rep. 2018;25(4):893-908 e7.

72. Seabright GE, Doores KJ, Burton DR, Crispin M. Protein and Glycan Mimicry in HIV Vaccine Design. J Mol Biol. 2019;431(12):2223-47.

73. Udenwobele DI, Su RC, Good SV, Ball TB, Varma Shrivastav S, Shrivastav A. Myristoylation: An Important Protein Modification in the Immune Response. Front Immunol. 2017;8:751.

74. Resh MD. Trafficking and signaling by fatty-acylated and prenylated proteins. Nat Chem Biol. 2006;2(11):584-90.

75. Maurer-Stroh S, Eisenhaber F. Myristoylation of viral and bacterial proteins. Trends Microbiol. 2004;12(4):178-85.

76. Hewitt EW, Lehner PJ. The ABC-transporter signature motif is required for peptide translocation but not peptide binding by TAP. Eur J Immunol. 2003;33(2):422-7.

77. Ding H, Guo M, Vidhyasagar V, Talwar T, Wu Y. The Q Motif Is Involved in DNA Binding but Not ATP Binding in ChlR1 Helicase. PLoS One. 2015;10(10):e0140755.

78. Crua Asensio N, Munoz Giner E, de Groot NS, Torrent Burgas M. Centrality in the host-pathogen interactome is associated with pathogen fitness during infection. Nat Commun. 2017;8:14092.

79. Ahmed H, Howton TC, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS. Network biology discovers pathogen contact points in host protein-protein interactomes. Nat Commun. 2018;9(1):2312.

Figure Legends

Figure 1a. Schematic of DLP: A host interactor protein (red) and pathogen protein (orange) that interacts with the same host protein (blue) share a similar domain X (yellow). In this way, the pathogen protein can mimic the domain of the host interactor protein and competes with it to bind to the host protein, thus causing the disease.**b. Schematic of MLP:** A host interactor protein (red) and pathogen protein (yellow) that interacts with the same host protein (blue) share a similar motif X (orange). In this way the pathogen protein can mimic the motif of the host interactor protein and competes with it to bind to the host protein, thus causing the disease.

Figure 2. Database schematic: The basic pipeline for search options is represented on the left and the basic workflow as well as the count of entities in the database are shown on the right.

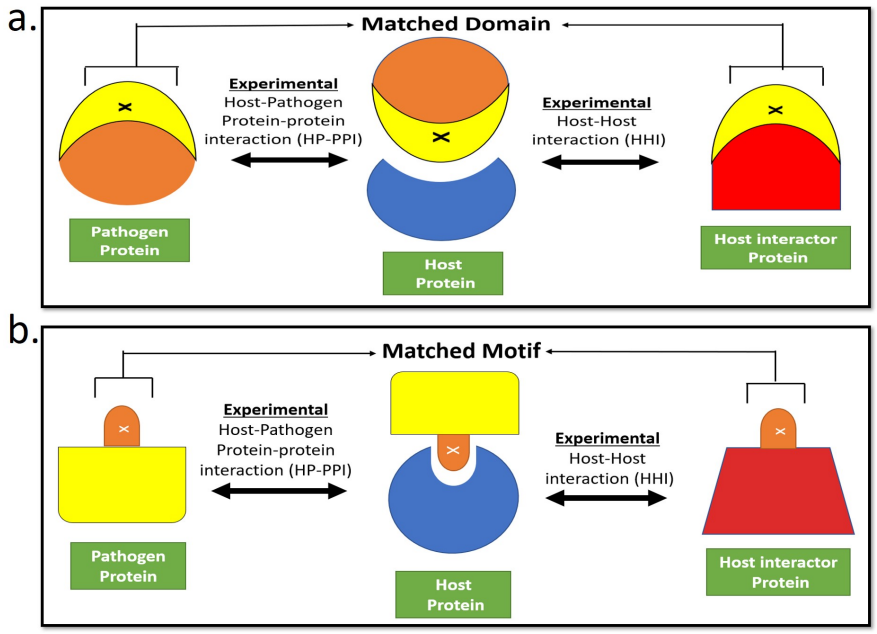
Figure 3. Global and local imitation of host proteins by pathogens : Graph depicting the total interactions and the interactions characterized by mimicked domains and motifs for different categories of pathogen.

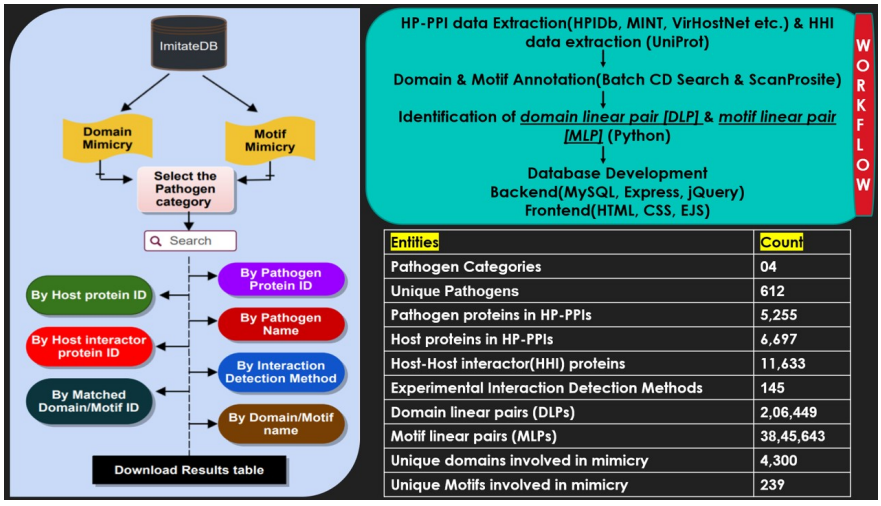
Figure 4 a. Frequently occurring domains: Bar graph showing the frequency of top 10 domains mimicked by pathogens in the database. The description of the domains is as follows: PHA03247- large tegument protein UL36; Smc- Chromosome segregation ATPase; SMC_prok_B- chromosome segregation protein SMC: common bacterial type; PKc- Catalytic domain of Protein Kinases; STKc_PknB_like- Catalytic domain of bacterial Serine/Threonine kinases, PknB and similar proteins; STKc_CMGC- Catalytic domain of CMGC family Serine/Threonine Kinases; STKc_CAMK- The catalytic domain of CAMK family Serine/Threonine Kinases; STKc_AMPK-like- Catalytic domain of AMP-activated protein kinase-like Serine/Threonine Kinases; STKc_PDK1- Catalytic domain of the Serine/Threonine Kinase, Phosphoinositide-dependent kinase 1; STKc_MLCK-like- Catalytic kinase domain of Myosin Light Chain Kinase-like Serine/Threonine Kinases. **b. Frequently occurring motifs:** Bar graph showing the frequency of top 10 motifs mimicked by pathogens in the database. The description of the motifs is as follows: PKC_PHOSPHO_SITE- Protein

kinase C phosphorylation site; CK2_PHOPHO_SITE- Casein kinase II phosphorylation site; MYRISTYL- N-myristoylation site; ASN_GLYCOSYLATION- N-glycosylation site; CAMP_PHOSPHO_SITE- cAMP- and cGMP-dependent protein kinase phosphorylation site; AMIDATION- Amidation site; TYR_PHOSPHO_-SITE_1- Tyrosine kinase phosphorylation site 1; TYR_PHOSPHO_SITE_2- Tyrosine kinase phosphorylation site 2; RGD- Cell attachment sequence; PRO_RICH- Proline-rich region profile.

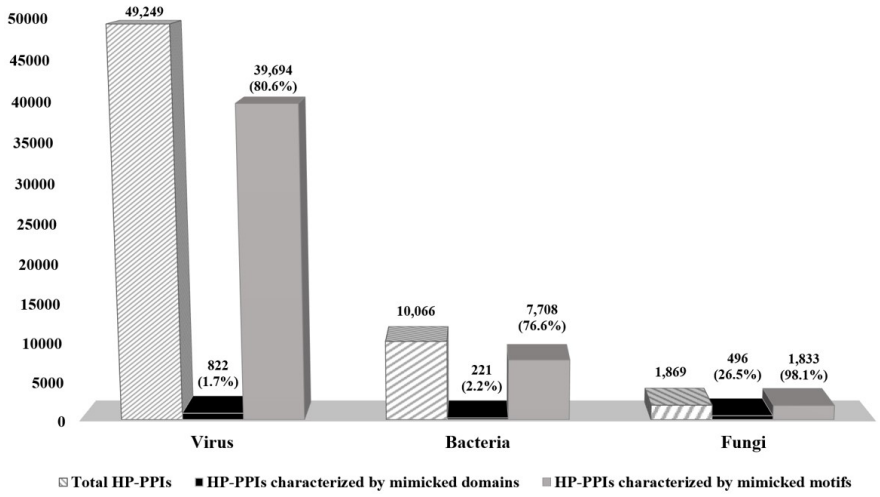
Figure 5 Enriched pathways in host proteins: Bar graph depicting the enriched pathways of host proteins in the database along with their enrichment percentage.

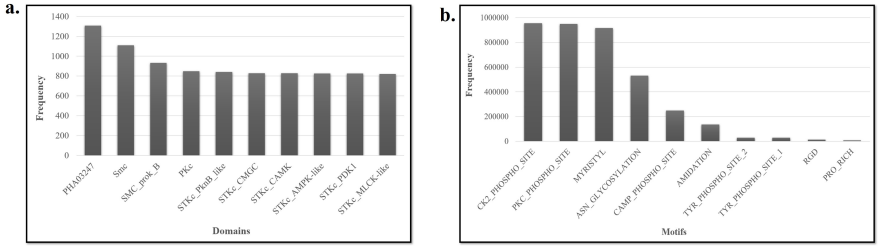
Figure 6 a. The ImitateDB web interface: Expanded view of the search panel of the web interface showing the steps to query the ImitateDB database. **b.Receive large result files by email:** Expanded view of the mailer popped up the on the ImitateDB interface.

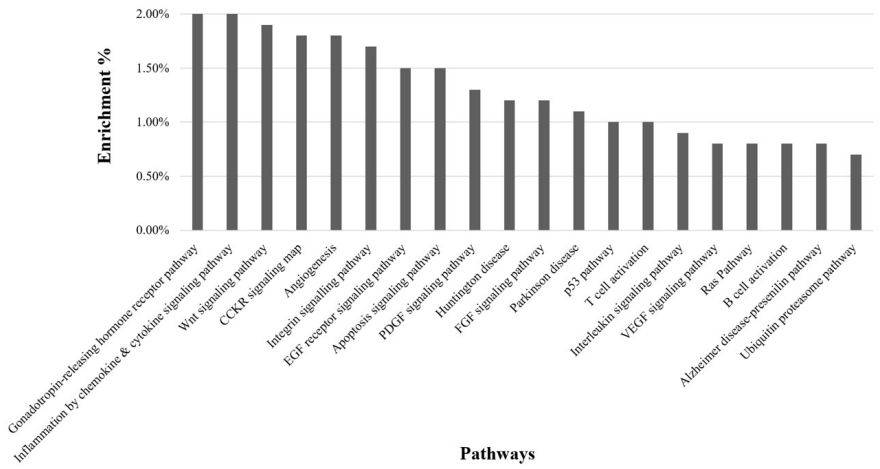


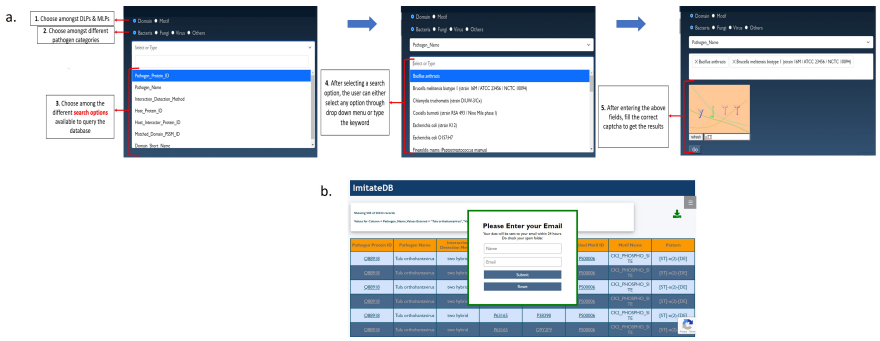


Entities	Count
Pathogen Categories	04
Unique Pathogens	612
Pathogen proteins in HP-PPIs	5,255
Host proteins in HP-PPIs	6,697
Host-Host interactor(HHI) proteins	11,633
Experimental Interaction Detection Methods	145
Domain linear pairs (DLPs)	2,06,449
Motif linear pairs (MLPs)	38,45,643
Unique domains involved in mimicry	4,300
Unique Motifs involved in mimicry	239









Hosted file

Table 1.docx available at https://authorea.com/users/426972/articles/531310-imitatedb-a-resource-for-domain-and-motif-mimicry-in-host-and-pathogen-proteins

Hosted file

Table 2.docx available at https://authorea.com/users/426972/articles/531310-imitatedb-a-resource-for-domain-and-motif-mimicry-in-host-and-pathogen-proteins

Hosted file

Table 3.docx available at https://authorea.com/users/426972/articles/531310-imitatedb-a-resource-for-domain-and-motif-mimicry-in-host-and-pathogen-proteins

Hosted file

Table 4.docx available at https://authorea.com/users/426972/articles/531310-imitatedb-a-resource-for-domain-and-motif-mimicry-in-host-and-pathogen-proteins

Hosted file

Table 5.docx available at https://authorea.com/users/426972/articles/531310-imitatedb-a-resource-for-domain-and-motif-mimicry-in-host-and-pathogen-proteins

Hosted file

Table 6.docx available at https://authorea.com/users/426972/articles/531310-imitatedb-a-resource-for-domain-and-motif-mimicry-in-host-and-pathogen-proteins

Hosted file

Table 7.docx available at https://authorea.com/users/426972/articles/531310-imitatedb-a-resource-for-domain-and-motif-mimicry-in-host-and-pathogen-proteins

Hosted file

Table 8.docx available at https://authorea.com/users/426972/articles/531310-imitatedb-a-resource-for-domain-and-motif-mimicry-in-host-and-pathogen-proteins

Hosted file

Table 9.docx available at https://authorea.com/users/426972/articles/531310-imitatedb-a-resource-for-domain-and-motif-mimicry-in-host-and-pathogen-proteins