


# SCIENTIFIC REPORTS



OPEN

## MorCVD: A Unified Database for Host-Pathogen Protein-Protein Interactions of Cardiovascular Diseases Related to Microbes

Nirupma Singh<sup>1</sup>, Venugopal Bhatia<sup>1</sup>, Shubham Singh<sup>2</sup> & Sonika Bhatnagar<sup>1</sup> 

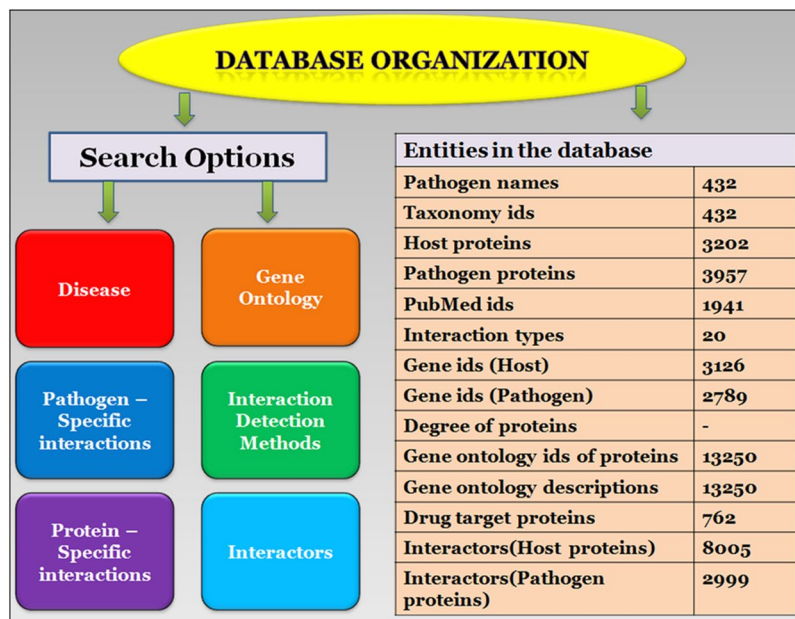
Microbe induced cardiovascular diseases (CVDs) are less studied at present. Host-pathogen interactions (HPIs) between human proteins and microbial proteins associated with CVD can be found dispersed in existing molecular interaction databases. MorCVD database is a curated resource that combines 23,377 protein interactions between human host and 432 unique pathogens involved in CVDs in a single intuitive web application. It covers endocarditis, myocarditis, pericarditis and 16 other microbe induced CVDs. The HPI information has been compiled, curated, and presented in a freely accessible web interface (<http://morcvd.sblab-nsit.net/About>). Apart from organization, enrichment of the HPI data was done by adding hyperlinked protein ID, PubMed, gene ontology records. For each protein in the database, drug target and interactors (same as well as different species) information has been provided. The database can be searched by disease, protein ID, pathogen name or interaction detection method. Interactions detected by more than one method can also be listed. The information can be presented in tabular form or downloaded. A comprehensive help file has been developed to explain the various options available. Hence, MorCVD acts as a unified resource for retrieval of HPI data for researchers in CVD and microbiology.

Cardiovascular diseases (CVDs) are amongst the most common cause of mortality and account for high morbidity across the globe<sup>1,2</sup>. Some of the major cardiovascular diseases include cardiac hypertrophy, rheumatic heart disease, ischemic heart disease, coronary artery disease, peripheral artery disease, and cerebrovascular disease<sup>3</sup>. In the past few years, the paradigm that microorganisms play an important role in the initiation and progression of CVDs has emerged. This paradigm has been supported by multiple epidemiological studies that have established positive associations between the risk of cardiovascular disease and markers of infection. Evidence implicating the infection by microbes in CVD includes the identification of viruses and bacteria in atherosclerotic plaques<sup>4</sup>, sero-epidemiological data<sup>5</sup>, and a strong association between specific infections such as Cytomegalovirus with transplant atherosclerosis<sup>6,7</sup>.

Common cardiovascular diseases caused by infection with microorganisms are endocarditis, pericarditis and myocarditis<sup>8</sup>. Infectious organisms or their structural components show the ability to induce proatherogenic and prothrombotic responses in cells relevant to atherogenesis (smooth muscle cells, monocyte-macrophages, T-cells, and endothelial cells)<sup>9</sup>. Microbial species that are found to be present in CVD affected patient samples include *Chlamydia pneumoniae*, *Porphyromonas gingivalis*, *Helicobacter pylori*, Influenza virus, Hepatitis C virus, Cytomegalovirus, Human Immunodeficiency Virus, Coxsackie Virus, and *Staphylococcus* species<sup>10</sup>. The mechanism of interaction of these microbial species with the human system at the molecular level and their involvement in the initiation, progression and severity of CVDs is yet to be elucidated.

Currently available CVD related databases like CardioGenBase database<sup>11</sup>, CADgene database<sup>12</sup> provide molecular and protein-protein interactions (PPIs) information but do not cover any HPI information of CVDs caused by microorganisms. Several databases list HPI data at the level of interacting proteins e.g. Reactome<sup>13</sup>, HMDAD<sup>14</sup>, PHI-base<sup>15</sup>, VirusMentha<sup>16</sup>, OrthoHPI<sup>17</sup>, VirusMINT<sup>18</sup>, EHFPI<sup>19</sup>, MatrixDB<sup>20</sup>, BioGrid<sup>21</sup>, HPIDb<sup>22</sup>,

<sup>1</sup>Computational and Structural Biology Laboratory, Division of Biological Sciences and Engineering, Netaji Subhas University of Technology, Dwarka, New Delhi, 110078, India. <sup>2</sup>Division of Computer Engineering, Netaji Subhas University of Technology, Dwarka, New Delhi, 110078, India. Venugopal Bhatia and Shubham Singh contributed equally. Correspondence and requests for materials should be addressed to S.B. (email: [sbhatnagar@nsut.ac.in](mailto:sbhatnagar@nsut.ac.in))



**Figure 1.** A schematic of database organization. Search options of the database are shown on the left and count of occurrence of each entity in the data is listed on the right.

MINT<sup>23</sup>, IMEx<sup>24</sup>, IntAct<sup>25</sup>, UniProt<sup>26</sup>, MPIDB<sup>27</sup>, VirHostNet<sup>28</sup>, I2D<sup>29</sup>, InnateDB<sup>30</sup>, DIP<sup>31</sup>, Mentha<sup>32</sup> and PHISTO<sup>33</sup>. Of these, only BioGrid, HPIDb, MINT, IntAct, UniProt, MPIDB, VirHostNet, I2D, MatrixDB, InnateDB and DIP contain limited and scattered information of HPIs leading to CVDs.

At present, independently collected host pathogen protein interaction data in microbe induced CVD is housed in various databases. This poses a big problem since the interactions across all these databases are repeated, scattered or highly fragmented. At present, no database is available that comprehensively lists all the unique protein interactions between host and pathogen in CVDs in a standard, enriched format. A researcher requiring such data has to first take the pains of aggregating the data from various databases available online. The data then has to be filtered, cleaned, processed and verified before finally being used. Therefore, we have developed a new database named MorCVD solely dedicated to the information comprising the interactions between proteins of human and microbial species leading to different types of CVD.

The keywords pertaining to microbe induced CVDs were finalized initially and a list of genes associated with those keywords data was collected from relevant databases. The HPIs corresponding to the genes were mined separately from twelve different databases, cleaned and enriched. For each interaction, gene ontologies, drug target data and interactors are available in MorCVD. The protein, literature and ontology records have been integrated through hyperlinks. The MorCVD MySQL database has a web interface developed using HTML, asp.net framework, CSS, JQuery, JavaScript and Microsoft Visual Studio. Several search options were developed to allow query of the database namely “Disease”, “Pathogen-Specific Interactions”, “Protein-Specific Interactions”, “Gene Ontologies”, “Interaction Detection Methods” and Interactors and Drug Targets”. MorCVD database is freely accessible at <http://morcvd.sblab-nsit.net/About> and will act as a unique resource for researchers in the field of microbiology and CVD.

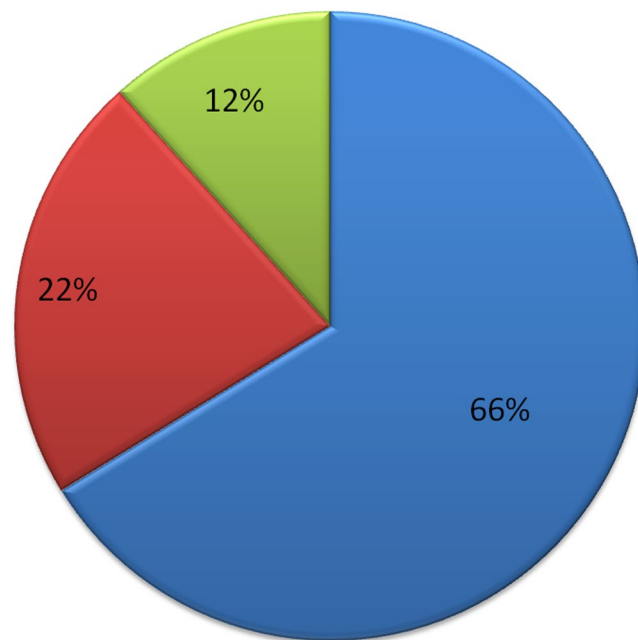
## Results and Discussion

The MorCVD database provides comprehensive information on human and pathogen proteins involved in the interactions leading to CVDs. The web application is compatible with multiple browsers like Chrome, Internet Explorer, Firefox, Safari and Microsoft Edge.

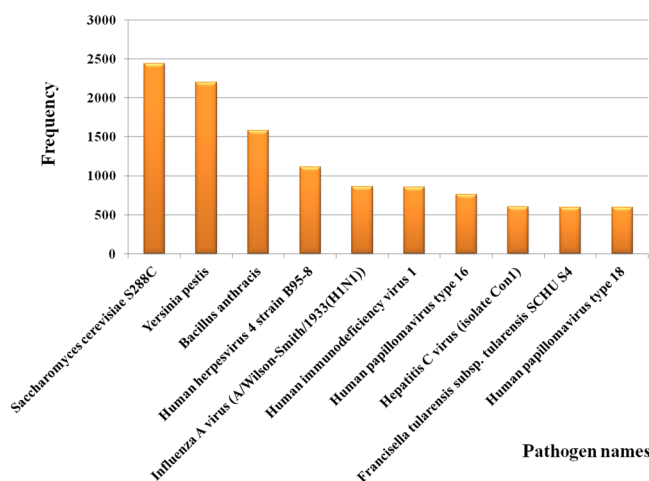
After processing the data collected from different databases, 23,377 unique HPIs were obtained. The associated data fields of the database were pathogen name, taxon category, taxonomy id, UniProt accession number, UniProt entry name, gene symbol, gene id, gene ontology, secondary interactors, drug target information of host and pathogen proteins, types of interaction, interaction detection methods, and confidence score of the interaction along with relevant PubMed IDs. The data fields like protein IDs, Pubmed IDs and gene ontology records were associated with corresponding clickable hyperlinks.

The confidence scores are assigned to host pathogen interactions to have a metric for gauging the likelihood of an interaction being biologically significant. These scores combine a variety of scoring schemes based on the number of publications supporting the interaction, number of detection methods, interaction types, shared gene ontologies etc. The confidence score for each HPI in MorCVD has been taken from the respective database from which the interaction data was originally obtained. The database was organized according to the broad schema shown in Fig. 1. The frequency of different types of pathogen species is shown in the form of a pie chart in Fig. 2.

The top 10 most frequently occurring pathogens, pathogen proteins and host proteins in the database along with their frequencies are shown in Figs 3, 4 and 5 respectively. The number of interactions that were found for



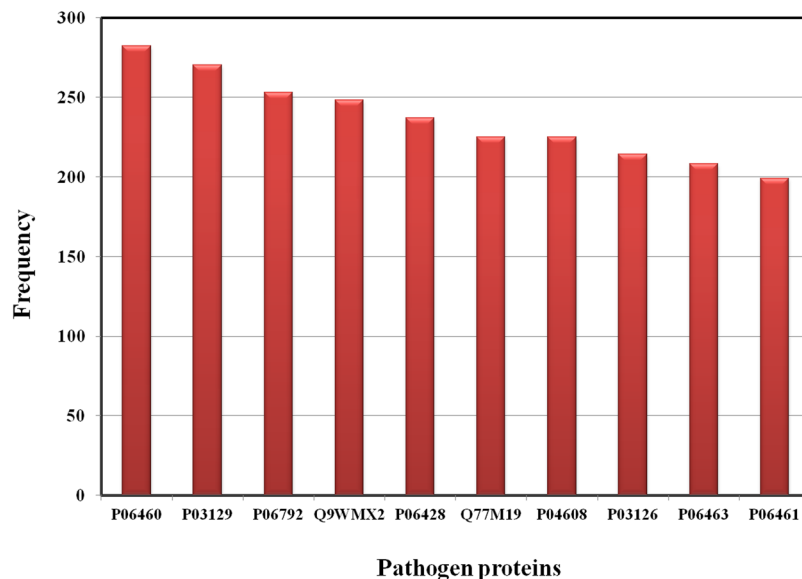
**Figure 2.** Pie chart showing the frequency of different types of pathogens in the database. Red denotes bacterial species; blue represents virus and green shows fungi as well as other pathogen species.



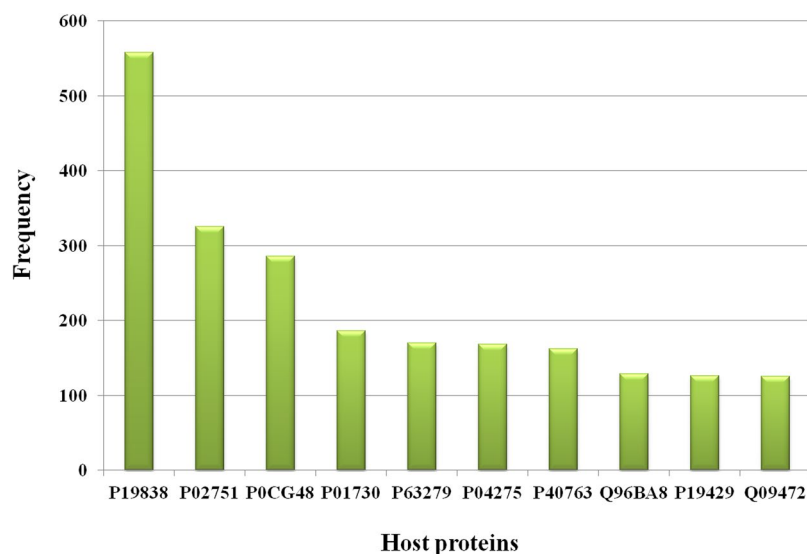
**Figure 3.** Frequency distribution of pathogens. The bar graph shows the number of top 10 pathogens in the database.

each of the disease term searched is shown in Table 1. Among all the interactions the disease ‘viral myocarditis’ was found to have the maximum number of interactions in the database. The number of HPIs for each type of pathogen involved in microbial CVDs (Table 2) and their frequency by class (Fig. 2) showed that till date virus-host interactions have been most widely studied in CVDs. Among viral species, predominantly occurring viruses were Human herpesvirus 4 strain B95-8, Influenza A virus (A/Wilson-Smith/1933(H1N1)), Human immunodeficiency virus 1, Human papillomavirus type 16 and Hepatitis C virus (isolate Con1). In case of bacterial species, the predominant microorganisms were *Saccharomyces cerevisiae* S288C, *Yersinia pestis*, *Bacillus Anthracis* and *Francisella tularensis* subsp. *tularensis* SCHU S4.

The enrichment of initial entities was further done for better analysis of the data. The parameters considered for the enrichment were gene ontologies, drug target information and interactors (same and different species) of host and pathogen proteins. Gene ontology is an important parameter to consider for enrichment analysis because it provides the information about biological process, molecular function and cellular component of the proteins<sup>34</sup>. By considering these three aspects of gene ontologies we get an idea about overrepresented or underrepresented functions and processes of proteins in our data as well as the location of proteins in a particular disease condition. The pathogen first has to make contact with the host via either extracellular secreted or membrane host proteins and further PPIs take place in the cytoplasmic environment. Therefore, the proteins were found equally enriched



**Figure 4.** Frequency distribution of pathogen proteins. The bar graph shows the number of top 10 pathogen proteins in the database. The description of UniProt accession number for pathogen proteins is as follows: P06460 - Probable protein E5A(Human papillomavirus type 6b), P03129 - Protein E7 (Human papillomavirus type 16), P06792 - Probable protein E5 (Human papillomavirus type 18), Q9WMX2 - Genome polyprotein (Hepatitis C virus genotype 1b (isolate Con1) (HCV)), P06428 - Protein E6 (Human papillomavirus type 8), Q77M19 - V protein (Measles virus (strain Edmonston-Schwarz vaccine) (MeV)), P04608 - Protein Tat (Human immunodeficiency virus type 1 group M subtype B (isolate HXB2) (HIV-1)), P03126 - Protein E6 (Human papillomavirus type 16), P06463 - Protein E6 (Human papillomavirus type 18), P06461 - Probable protein E5B (Human papillomavirus type 6b).



**Figure 5.** Frequency distribution of host proteins. The bar graph shows the number of top 10 host proteins in the database. The description of UniProt accession number for human host proteins is as follows: P19838 - Nuclear factor NF-kappa-B p105 subunit; P02751 - Fibronectin; P0CG48 - Polyubiquitin-C; P01730 - T-cell surface glycoprotein CD4; P63279 - SUMO-conjugating enzyme UBC9; P04275 - von Willebrand factor; P40763 - Signal transducer and activator of transcription 3; Q96BA8 - Cyclic AMP-responsive element-binding protein 3-like protein 1; P19429 - Troponin I, cardiac muscle; Q09472 - Histone acetyltransferase p300.

in all the cellular compartments like cytoplasm (GO:0005737), membrane (GO:0016020) and integral component of membrane (GO:0016021) and were further scanned for enriched biological process and molecular function. The gene ontology molecular functions like protein binding(GO:0005515) and metal ion binding (GO:0046872) were found to be most enriched in the data both in case of human host as well as the pathogen. Some other gene

S. No.	Disease name	Number of interactions
1.	Viral Myocarditis	8002
2.	Dilated Cardiomyopathy	6155
3.	Endocarditis	3541
4.	Cardiovascular Infections	2786
5.	Hypereosinophilic Syndrome	749
6.	Pericarditis	675
7.	Myocarditis	495
8.	Bacterial Endocarditis	345
9.	Infective Endocarditis	273
10.	Peripartum Cardiomyopathy	100
11.	Native Valve Endocarditis	78
12.	Subacute Bacterial Endocarditis	53
13.	Acute Myocarditis	42
14.	<i>Staphylococcus aureus</i> Endocarditis	34
15.	Viral Cardiomyopathy	17
16.	Q-fever Endocarditis	16
17.	Chronic Myocarditis	8
18.	Endocarditis of mitral valve	5
19.	Camptodactyly-Arthropathy-Coxa Vara-Pericarditis Syndrome	3

**Table 1.** Number of interactions for each disease term.

S.No.	Pathogen	Number
1.	Virus	15514
2.	Bacteria	5184
3.	Fungi	2668
4.	Protozoa	24
5.	Amoebozoa	9
6.	Archaea	2

**Table 2.** Types of pathogen involved with number of interactions.

ontologies those were found to be predominantly occurring exclusively in case of pathogen proteins were molecular functions like nucleotide binding (GO:0000166) and ATP binding (GO:0005524). The predominantly occurring gene ontology biological process in case of host proteins was regulation of transcription (GO:0006355) while viral process (GO:0016032) in case of pathogen proteins. The top ten gene ontologies for host and pathogen proteins are listed in Table 3 and their frequencies are represented in the form of a pie chart in Figs 6 and 7 respectively.

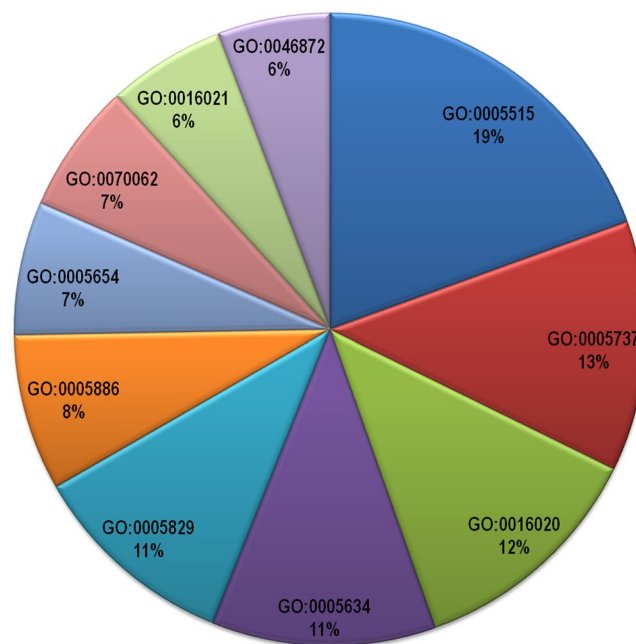
After gene ontology enrichment analysis, drug target analysis was done. A search for those proteins in the data identified 702 human proteins and 60 pathogen proteins as previously known drug targets. Drug target analysis was done to look for gene ontology pattern of drug targets and their interactions in the data. The frequently occurring gene ontologies for both human and pathogen drug target proteins were found to be molecular functions like protein binding (GO:0005515) and metal ion binding (GO:0046872). The prominent gene ontology biological processes in case of pathogen drug targets were the viral process (GO:0016032) and proteolysis (GO:0006508). The molecular functions more prominent exclusively in case of pathogen drug target proteins were hydrolase activity (GO:0016787) and transferase activity (GO:0016740).

The quantitative analysis done for the number of interactions between host and pathogen proteins provided a list of high degree proteins. It is important to consider the highly interacting host proteins because they represent the most influential entities in the data and these are the proteins approached most by the pathogen proteins. We found 30 host proteins and 69 pathogen proteins in our data that were high degree. The top 5 high degree host and pathogen proteins in the data are listed in Table 4. Amongst the host drug target proteins, the ones having the highest degree were Nuclear factor NF-kappa-B p105 subunit protein, Fibronectin, Polyubiquitin-C, T-cell surface glycoprotein CD4 and von Willebrand factor while in case of pathogen they were Genome polyprotein (Hepatitis C virus genotype 1b (isolate Con1) (HCV)), Genome polyprotein (Hepatitis C virus genotype 1a (isolate H) (HCV)) and Pol polyprotein (Human immunodeficiency virus 1).

In the last part of the analysis same and different species interactors for both host and pathogen proteins were added. It is important to study about all types of interactors because they together regulate a variety of cellular functions, including cell cycle progression, signal transduction, and metabolic pathways inside the body<sup>35</sup>. The proteins having a very large number of interactors are known to be essential as they play a key role in the

S. No.	Gene Ontology (human)	Gene Ontology (pathogen)
1.	protein binding (GO:0005515)	protein binding (GO:0005515)
2.	cytoplasm (GO:0005737)	membrane (GO:0016020)
3.	membrane (GO:0016020)	integral component of membrane (GO:0016021)
4.	nucleus (GO:0005634)	cytoplasm (GO:0005737)
5.	cytosol (GO:0005829)	metal ion binding (GO:0046872)
6.	plasma membrane (GO:0005886)	nucleotide binding (GO:0000166)
7.	nucleoplasm (GO:0005654)	viral process (GO:0016032)
8.	extracellular exosome (GO:0070062)	ATP binding (GO:0005524)
9.	integral component of membrane (GO:0016021)	transferase activity (GO:0016740)
10.	metal ion binding (GO:0046872)	hydrolase activity (GO:0016787)

**Table 3.** Predominant occurring gene ontology ids (top 10).



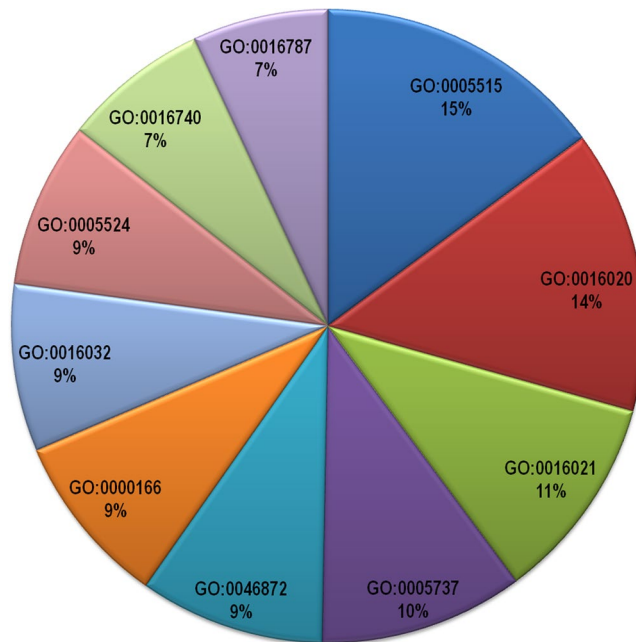
**Figure 6.** Pie Chart showing frequency of top 10 most prominent gene ontologies of host proteins in the database. The gene ontology description is as follows: GO:0005515 -protein binding, GO:0005737 - cytoplasm, GO:0016020 - membrane, GO:0005634 - nucleus, GO:0005829 - cytosol, GO:0005886 - plasma membrane, GO:0005654 - nucleoplasm, GO:0070062 - extracellular exosome, GO:0016021 - integral component of membrane, GO:0046872 - metal ion binding.

metabolism of the body and can disturb the functioning of the body if troubled. In our data, the human and pathogen proteins with maximum number of interactors are listed in Table 5.

After all the MySQL work done for database construction and analysis, a web interface named MorCVD database was developed for this data. It comprises several search options that fulfill the demand for extraction of HPI data and its enrichment parameters. They are as follows:

- Disease**  
 This option can be used to list all the HPIs related to a particular disease term. It requires the selection of a specific term from the drop-down menu of 19 disease terms displayed on the webpage. The user can choose any one of the disease terms to retrieve the desired data. The search result leads to the record of alphabetically ordered pathogen names, pathogen taxonomy identifiers, UniProt accession numbers, gene symbols and UniProt entry names, of host and pathogen proteins, a sortable confidence score, source database for the interaction and respective PubMed reference of the HPI. For every disease, results pertaining to specific host or pathogen protein can be found using the “Search by protein IDs” tab present on the result page. The data download option makes it convenient for the user to explore large number of results simultaneously.
- Pathogen-specific interactions**  
 This option allows the user to look for HPIs with respect to a particular pathogen query. It comprises a pathogen tab that lists all the pathogen names in a drop-down menu. When a pathogen is selected all the HPIs





**Figure 7.** Pie Chart showing frequency of top 10 most prominent gene ontologies of pathogen proteins in the database. The gene ontology description is as follows: GO: 0005515 - protein binding, GO:0016020 - membrane, GO:0016021 - integral component of membrane, GO:0005737 - cytoplasm, GO:0046872 - metal ion binding, GO:0000166 - nucleotide binding, GO:0016032 - viral process, GO:0005524 - ATP binding, GO:0016740 - transferase activity, GO:0016787 - hydrolase activity.

S. No.	Host proteins	Pathogen proteins
1.	Nuclear factor NF-kappa-B p105 subunit	Probable protein E5A (Human papillomavirus type 6b)
2.	Fibronectin	Protein E7 (Human papillomavirus type 16)
3.	Polyubiquitin-C	Probable protein E5 (Human papillomavirus type 18)
4.	T-cell surface glycoprotein CD4	Genome polyprotein (Hepatitis C virus genotype 1b (isolate Con1) (HCV))
5.	SUMO-conjugating enzyme UBC9	Protein E6 (Human papillomavirus type 8)

**Table 4.** High degree (highly interacting) proteins of data (top 5).

related to that particular pathogen are listed. This could help in determining the list of proteins interacting in the data for a particular pathogen.

- Protein-specific interactions**  
 This option allows the querying of the database by a particular protein. After the selection of host (i.e. either human host or pathogen-host), the option requires the input of respective protein's UniProt accession number. It then lists all the interactions related to that particular protein along with their gene identifiers, degrees of the listed proteins and the corresponding disease. In case of selecting pathogen as host, the user further needs to select a pathogen from the drop-down menu and then provide the UniProt accession number of protein for that particular pathogen.
- Gene Ontologies**  
 This option can be used to provide the gene ontology information of host and pathogen proteins present in the data. The user needs to provide UniProt accession numbers of either human or pathogen proteins in order to list the corresponding gene ontology identifiers with their class, description and quantitative value (on the basis of its occurrence in the data) of each gene ontology id.
- Interaction - detection methods**  
 This option lists the HPIs in the data by a particular interaction detection method. After selection of an interaction detection method from the drop-down menu, it lists all the HPIs that were identified by that particular interaction detection method along with interaction type between them. With the help of this option, we also get information about the type of interactions fetched out from a particular interaction detection method.  
 A separate page provides the option to search the interactions detected by multiple methods. The user has to select the number of methods from the drop down menu to obtain the interactions along with the respective interaction detection methods.

S. No.	Host proteins	Pathogen proteins
1.	Homeobox protein MOX-2	Ribosome-associated molecular chaperone SSB1 ( <i>Saccharomyces cerevisiae</i> (strain ATCC 204508/S288c) (Baker's yeast))
2.	Microtubule-associated tumor suppressor candidate 2	Alpha N-terminal protein methyltransferase 1 ( <i>Saccharomyces cerevisiae</i> (strain ATCC 204508/S288c) (Baker's yeast))
3.	MyoD family inhibitor	Translation initiation factor eIF-2B subunit beta ( <i>Saccharomyces cerevisiae</i> (strain ATCC 204508/S288c) (Baker's yeast))
4.	Proto-oncogene c-Rel	ATP-dependent molecular chaperone HSC82 ( <i>Saccharomyces cerevisiae</i> (strain ATCC 204508/S288c) (Baker's yeast))
5.	Cellular tumor antigen p53	Heat shock protein SSA1 ( <i>Saccharomyces cerevisiae</i> (strain ATCC 204508/S288c) (Baker's yeast))

**Table 5.** Proteins with maximum number of interactors (same and different species) (top 5).

#### • Interactors and Drug targets

This option provides the list of interactors of the human and pathogen proteins. The user can list the interactors of a particular protein by choosing a host and then providing the UniProt accession number of either the human or pathogen protein as input. This option also provides the drug target information of the queried protein, whether it is a drug target or not. The user can also look for common interactors between two proteins.

Hence, MorCVD is a database constructed to provide a comprehensive view of molecular information of HPis leading to CVDs. MorCVD encompasses a broader spectrum of data and provides us a unique resource that collates all the HPis involved in CVDs by integrating the information from biological databases present in a scattered manner. By providing effective search and browsing features, it operates as a flexible and user-friendly platform for the molecular study of microbial CVDs. It contains the gene ontology parameters, drug target information and interactors within the same species as well as different species of host and pathogen proteins. Comprehensive documentation of the database is available to the users through the Documentation option (Supplementary File 1). A link to a page containing a brief description of all the 116 detection methods for determining the interactions has been included in the Documentation (Supplementary Table S1 in Supplementary File 1).

MorCVD database gives a collective idea about biochemical and physiological properties of proteins. Analysis of the data provided the information that viruses are more likely to be involved in case of microbial CVDs as compared to other pathogen species as determined by the frequency of viruses in the HPI data. MorCVD lists enriched HPI data relevant to CVDs and is sufficient to show that most of the interactions have been determined for virus proteins and many more bacterial HPis need to be determined. It also provides the information about already known drug target proteins present in our data. More number of host proteins were identified as drug targets in the data as compared to pathogen proteins. Thus, MorCVD is a unified database which provides the information about HPis involved in microbial CVDs, their functional features, and some biological properties.

**Methods.** The building of the database included the following steps:

#### 1. Database mining and curation

The information for the database was collected by defining a list of search terms through a literature survey. The infection of microorganisms in the heart is classified mainly into three types of cardiovascular inflammatory disorders namely 'endocarditis', 'myocarditis' and 'pericarditis'<sup>8,36</sup>. Consecutively, some other terms were also found from DisGeNet database<sup>37</sup> associated with former three main disease terms and were cross-checked from the literature. Those terms were 'dilated cardiomyopathy', 'viral cardiomyopathy', 'viral myocarditis', 'acute myocarditis', 'chronic myocarditis', 'peripartum cardiomyopathy', 'Camptodactyly-Arthropathy-Coxa Vara-Pericarditis Syndrome', 'hypereosinophilic syndrome', 'bacterial endocarditis', 'infective endocarditis', 'Q – fever endocarditis', 'subacute bacterial endocarditis', '*Staphylococcus aureus* endocarditis', 'native valve endocarditis', 'endocarditis of mitral valve' and 'cardiovascular infections'. After defining the list of search terms, genes related to them were extracted from the databases, namely OpenTargets<sup>38</sup> and DisGeNet.

The combined gene list of both the databases after removing duplicates was fed into the following databases i.e. BioGrid, HPIDb, MINT, IntAct, UniProt, MPIDB, VirHostNet, I2D, MatrixDB, InnateDB and DIP in order to extract the relevant HPI information. HPI data collected from these resources included the corresponding UniProt accession numbers, UniProt entry names and gene symbols for interacting proteins of human host and pathogen, interaction detection method, confidence score of the interaction, interaction type between the two proteins, pathogen names for pathogen proteins with their taxon id and taxon category and PubMed id reference of the HPI. There were also some HPI databases that did not provide any relevant data for our objective, namely Reactome, HMDAD, PHI-base, VirusMentha, OrthoHPI, VirusMINT and EHFPI.

#### 2. Data processing

The raw data collected was filtered to ensure that the relevant data exclusively dedicated to pathogen protein interactions with proteins of the human cardiovascular system was retained. A UniProt accession number was used to identify the protein molecules uniformly that were collected from different sources. The pathogen names were also checked for differences in syntax/nomenclature and were transformed into



the single uniform format on the basis of same UniProt Taxon identifier. Duplicate records were removed from the data to prevent redundancy.

### 3. Data enrichment

The data was further processed and enriched with additional parameters. A list of unique host and pathogen proteins was extracted from the initial HPI data. Gene id numbers were added to this protein molecules list using db2db tool of the bioDBnet database. The unique protein entities of both human and pathogens were also enriched with information about their gene ontologies. The same species and different species interactors of host and pathogen proteins obtained from the UniProt database were added and also the drug target information obtained from DrugBank 3.0<sup>39</sup>. An indicator of the number of interactions (degree) was assigned to all the unique host proteins and pathogen proteins using MySQL command line. The Gene Ontology (GO) identifiers were also given a quantitative value on the basis of their occurrence in the finally processed data.

### 4. Database development

After processing and gathering of final data, MySQL relational database management system (RDMS) was used to construct a database by making use of various MySQL tools. The data was uploaded on the MySQL server localhost using MySQL workbench and query commands were made in the Command line Client of MySQL server. Several constraints like primary keys and foreign keys were assigned to several entities of tables in the database to remove and prevent further redundancy and ambiguities from the data in order to make a normalized database. Next, we proceeded to deploy our database in the form of a website. For this purpose, we used asp.net tools (C#) to develop the backend of the website environment and to link the MySQL database. The front end of the website was developed using HTML, CSS, JQuery and JavaScript. The use of Microsoft Visual Studio was also made to connect the asp.net backend environment with the HTML script front end so as to develop the whole web application. The web interface includes several search options for specific query and retrieval of data pertaining to 'Disease', 'Pathogen-Specific Interactions', 'Protein-Specific Interactions', 'Gene Ontologies', 'Interaction Detection Methods' and 'Interactors and Drug Targets'. The web interface also has a 'Contact us' page which includes all the documentation about the database and also a query form for the submission of any type queries or bug reports by the user.

### 5. Data analysis

Analysis of the database was done using R studio, MySQL and Microsoft Excel. R script was used to sort the entities and quantitatively measure the occurrence of pathogens, pathogen proteins, pathogen taxon categories, and host proteins, interactions per disease term, gene ontologies, detection methods, interaction types, gene ids and the number of interactors of host and pathogen proteins in the data. R script was also used to perform drug target analysis to look for the number of interactions of drug target proteins in the data and their gene ontology pattern. The graphs showing the frequency of the top ten maximally occurring entities of the database were made using Microsoft Excel. Quantitative analysis for the number of interactions of host and pathogen proteins was done in MySQL. A degree value was assigned to each unique protein (both human and pathogen) on the basis of the number of their interactions in the data. The high degree (highly interacting) proteins were obtained using a cutoff greater than three standard deviations from the mean.

## Data Availability

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

## References

- Mendis, S. & Norrving, B. Global atlas on cardiovascular disease prevention and control: World Health Organization, Vol. 924 (2011).
- Campbell, L. A. & Rosenfeld, M. E. Infection and Atherosclerosis Development. *Archives of medical research* **46**, 339–350 (2015).
- Larsen, T. R. *et al.* Lack of association between cystatin C and different coronary atherosclerotic manifestations. *Scandinavian Journal of Clinical and Laboratory Investigation* **77**, 574–581 (2017).
- Libby, P., Egan, D. & Skarlatos, S. Roles of infectious agents in atherosclerosis and restenosis: an assessment of the evidence and need for future research. *Circulation* **96**, 4095–4103 (1997).
- Shah, P. K. Plaque disruption and thrombosis. *Cardiology Clinics* **17**, 271–281 (1999).
- Fyfe, A. I. Transplant atherosclerosis: the clinical syndrome, pathogenesis and possible model of spontaneous atherosclerosis. *Can J Cardiol* **8**, 509–519 (1992).
- Nicholson, A. C. & Hajjar, D. P. Herpesviruses in Atherosclerosis and Thrombosis. *Arteriosclerosis, Thrombosis, and Vascular Biology* **18**, 339 (1998).
- Schöffel, N., Vitzthum, K., Mache, S., Groneberg, D. A. & Quarcoo, D. The Role of Endocarditis, Myocarditis and Pericarditis in Qualitative and Quantitative Data Analysis. *International Journal of Environmental Research and Public Health* **6**, 2919–2933 (2009).
- Gurfinkel, E. *et al.* Treatment with the antibiotic roxithromycin in patients with acute non-Q-wave coronary syndromes. The final report of the ROXIS Study. *Eur Heart J* **20**, 121–127 (1999).
- Ayada, K. *et al.* Chronic Infections and Atherosclerosis. *Annals of the New York Academy of Sciences* **1108**, 594–602 (2007).
- Alexandar, V. *et al.* CardioGenBase: A Literature Based Multi-Omics Database for Major Cardiovascular Diseases. *PLOS ONE* **10**, e0143188 (2015).
- Liu, H. *et al.* CADgene: a comprehensive database for coronary artery disease genes. *Nucleic Acids Research* **39**, D991–D996 (2011).
- Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* **39**, D691–D697 (2011).
- Ma, W. *et al.* An analysis of human microbe–disease associations. *Briefings in Bioinformatics* **18**, 85–97 (2017).
- Urban, M. *et al.* The Pathogen-Host Interactions database (PHI-base): additions and future developments. *Nucleic Acids Research* **43**, D645–D655 (2015).
- Calderone, A., Licata, L. & Cesareni, G. VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Research* **43**, D588–D592 (2015).
- Cuesta Astroz, Y., Santos, A., Oliveira, G. & Jensen, L. J. An integrative method to unravel the host-parasite interactome: An orthology-based approach. *bioRxiv* (2017).

18. Chatr-aryamontri, A. *et al.* VirusMINT: a viral protein interaction database. *Nucleic Acids Research* **37**, D669–D673 (2009).
19. Liu, Y. *et al.* EHFPI: a database and analysis resource of essential host factors for pathogenic infection. *Nucleic Acids Research* **43**, D946–D955 (2015).
20. 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 Research* **43**, D321–D327 (2015).
21. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34**, D535–D539 (2006).
22. Ammari, M. G., Gresham, C. R., McCarthy, F. M. & Nanduri, B. HPIDB 2.0: a curated database for host–pathogen interactions. *Database: The Journal of Biological Databases and Curation* **2016**, baw103 (2016).
23. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Research* **35**, D572–D574 (2007).
24. Orchard, S. *et al.* Protein Interaction Data Curation - The International Molecular Exchange Consortium (IMEx). *Nature methods* **9**, 345–350 (2012).
25. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Research* **32**, D452–D455 (2004).
26. The UniProt Consortium UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**, D158–D169 (2017).
27. Goll, J. *et al.* MPIDB: the microbial protein interaction database. *Bioinformatics* **24**, 1743–1744 (2008).
28. Navratil, V. *et al.* VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus–host interaction networks. *Nucleic Acids Research* **37**, D661–D668 (2009).
29. Brown, K. R. & Jurisica, I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biology* **8**, R95–R95 (2007).
30. Lynn, D. J. *et al.* InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Molecular Systems Biology* **4**, 218–218 (2008).
31. Xenarios, I. *et al.* DIP: the Database of Interacting Proteins. *Nucleic Acids Research* **28**, 289–291 (2000).
32. Calderone, A., Castagnoli, L. & Cesareni, G. mentha: a resource for browsing integrated protein–interaction networks. *Nature methods* **10**, 690 (2013).
33. Durmuş Tekir, S. *et al.* PHISTO: pathogen–host interaction search tool. *Bioinformatics* **29**, 1357–1358 (2013).
34. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29 (2000).
35. De Las Rivas, J. & Fontanillo, C. Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology* **6**, e1000807 (2010).
36. Murillo, H. *et al.* Infectious Diseases of the Heart: Pathophysiology, Clinical and Imaging Overview. *RadioGraphics* **36**, 963–983 (2016).
37. Piñero, J. *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database: The Journal of Biological Databases and Curation* **2015**, bav028 (2015).
38. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research* **45**, D985–D994 (2017).
39. Knox, C. *et al.* DrugBank 3.0: a comprehensive resource for ‘Omics’ research on drugs. *Nucleic Acids Research* **39**, D1035–D1041 (2011).

## Acknowledgements

NS acknowledges Council of Scientific and Industrial Research (CSIR) – JRF(09/836(0021)/2016-EMR-I) for financial support in research.

## Author Contributions

N.S. contributed in organizing the data, development of the database, wrote the main manuscript text and prepared figures. V.B. and S.S. contributed in developing the backend and front end of the database. All authors contributed in the analysis of the data. S.B. was involved in design and supervision of the study. All authors reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-019-40704-5>.

**Competing Interests:** The authors declare no competing interests.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019