ArVirInd—a database of arboviral antigenic proteins from the Indian subcontinent

Nitin Motilal Atre, Kalichamy Alagarasu and Pratip Shil

Bioinformatics, ICMR National Institute of Virology Pune, Pune, Maharashtra, India

ABSTRACT

Background. Studies on antigenic proteins for arboviruses are important for providing diagnostics and vaccine development. India and its neighboring countries have a huge burden of arboviral diseases. Data mining for country-specific sequences from existing bioinformatics databases is cumbersome and time-consuming. This necessitated the development of a database of antigenic proteins from arboviruses isolated from the countries of the Indian subcontinent.

Methods. Arboviral antigenic protein sequences were obtained from the NCBI and other databases. *In silico* antigenic characterization was performed (Epitope predictions) and data was incorporated into the database. The front end was designed and developed using HTML, CSS, and PHP. For the backend of the database, we have used MySQL.

Results. A database, named ArVirInd, is created as a repository of information on curated antigenic proteins. This enlists sequences by country and year of outbreak or origin of the viral strain. For each entry, antigenic information is provided along with functional sites, *etc*. Researchers can search this database by virus/protein name, country, and year of collection (or in combination) as well as peptide search for epitopes. It is available publicly via the Internet at http://www.arvirind.co.in. ArVirInd will be useful in the study of immune informatics, diagnostics, and vaccinology for arboviruses.

INTRODUCTION

With the effects of climate change becoming more prominent, the enhanced propagation of arthropods has resulted in the emergence and re-emergence of Arthropod-borne viruses (arboviruses) worldwide (*Ryan et al., 2017*; *Gould & Higgs, 2009*; *Mourya et al., 2016*). The densely populated South Asian region, geographically known as the Indian subcontinent, with its unique monsoon-driven climate systems, is particularly vulnerable to arboviral diseases such as dengue, chikungunya, Japanese encephalitis, West Nile, Crimean-Congo hemorrhagic fever, Kyasanur Forest disease, *etc*. are on the rise (*Nagpal et al., 2016*; *Dhiman et al., 2010*; *Shil, 2019*; *Mourya et al., 2016*). In India, between 2015 and 2021, Dengue affected 913,817 individuals (1,490 deaths), Chikungunya affected 82,429 individuals (0 deaths), and Japanease Encephalitis affected 11,283 (1,414 deaths) (NVBDCP

https://nvbdcp.gov.in/index.php). Apart from this, there were sporadic outbreaks of West Nile, and Zika. Dengue and Chikungunya are prevalent in all the countries of the Indian subcontinent (Bangladesh (*Haider et al., 2021*), Pakistan (*Islam et al., 2022*), Nepal (*Rijal et al., 2021*), *etc.*). In 2019 Nepal experienced huge epidemic of dengue covering 68 out of 77 districts (*Adhikari & Subedi, 2020*).

The re-emergence of these diseases is associated with complex factors such as viral recombination and mutation, leading to more virulent and adaptive strains (*Tabachnick, 2016*; *Baylis, 2017*). This necessitates analyses of genome and proteome information for tracing viral evolution and phylogenetics, identification of effects of mutations in newly emerging strains, and also the design and development of chimeric peptides for effective diagnostics and/or vaccines. All this research needs basic information about the nucleotide and protein sequence composition and characteristics from viral strains and *in silico* analyses using bioinformatics and immuno-informatics tools.

Since the last decade of the 20th-century emergence of bioinformatics tools, techniques and databases has revolutionized biological research. A huge amount of biological information generated from laboratory experiments is warehoused in databases, which are freely available through the internet. NCBI (https://www.ncbi.nlm.nih.gov/), Protein Information Resource (PIR) (https://proteininformationresource.org/), DDBJ (https://www.ddbj.nig.ac.jp/index-e.html), ExPASy (https://www.expasy.org/), Protein ligand interaction database (PLID) (https://ngdc.cncb.ac.cn/databasecommons/database/id/3794), TTRMDB (http://vit.ac.in/ttrmdb), and dbHDPLS (https://www.dlearningapp.com/web/dbDPLS/index.php) are a few examples of large and popular databases providing information on proteins, protein families and also provide users with online software tools for bioinformatics analyses (*Reddy et al., 2008*; *Zhu et al., 2019*; *Srinivasan, Natarajan & Rajasekaran, 2020*). Bioinformatics databases dedicated to viruses are available. Virus-Host DB is a database that provides users with information on various viral proteins (Virus-Host DB. URL: https://www.genome.jp/virushostdb/). The Virus Pathogen Database and Analyses resource (ViPR) is a valuable resource for viral proteins (https://www.viprbrc.org/). There also exists a database of flaviviruses, FLAVIdB (cvc.dfci.harvard.edu/flavi), which is of limited scope. The most recent dedicated database on dengue is DenvInt (*Dey & Mukhopadhyay, 2017*).

Bioinformatics plays important role in immunological studies and viral diagnostics like the design of ELISA based tests, chimeric peptides, antigen-antibody interactions, *etc.* (*Gangwar et al., 2011*; *Gangwar et al., 2012*; *Shil, Chavan & Cherian, 2011a*; *Shil, Chavan & Cherian, 2011b*; *Pavitrakar et al., 2018*; *Pavitrakar et al., 2020*) . Hence, quick and easy data retrieval from databases is essential. The practical difficulty faced while searching for Indian sequences is that none of the existing databases enlist entries by country of origin and date of sample collection or outbreak. Most databases enlist sequences by date of publication. So data mining from existing large databases is cumbersome as a large number of entries appears and each had to be read manually to identify the country of origin. Considering the huge burden of arboviral diseases in India and its immediate neighborhood, providing effective diagnostics and phylogenetic studies for tracing viral evolution are important

aspects of virology research. Hence, there is a need to develop a database dedicated to arboviral antigenic proteins from strains isolated from the Indian subcontinent.

In this article, we describe the development of a bioinformatics database, ArVirInd, dedicated to arboviral antigenic proteins covering strains reported from India and its neighboring countries. The database is being developed as a knowledge base and envisages providing information on antigenic characteristics apart from the sequence composition.

MATERIALS & METHODS

Data mining

The antigenic protein sequences for arboviruses isolated from the countries of the Indian sub-continent were retrieved from NCBI. We then searched the internet for the related publication for each sequence and traced out the date of collection or year of the outbreak and the country of origin and information on the host organism (mosquito or human, *etc.*). Searches were made for any available 3D structure for the sequence in PDB. Annotation has been provided for each database entry. The inclusion criterion included complete amino acid sequences for the concerned proteins. Exclusion criterion: all partial amino acid sequences were rejected.

Data analyses and annotation

For the flaviviruses such as Dengue, West Nile, and Japanese encephalitis the surface glycoprotein (E-protein) and non-structural protein1 (NS1) have been considered for inclusion in the database. For Chikungunya, belonging to the Togaviridae family, the surface proteins E1 and E2 have been considered. The database also includes Chandipura virus envelop protein (G-protein).

Each amino acid sequence was evaluated for antigenic potential using VaxiJen 2.0 server available at http:www.ddg-pharmac.net/vaxijen/VaxiJen/vaxiJen.html (*Doytchinova & Flower, 2007*). Then the sequences were considered for B-cell epitope prediction using different tools available at the Immune Epitope Database (IEDB) analyses resource (http://www.iedb.org/). Best results from the Kolaskar method of B-cell epitope prediction were considered for inclusion for each entry. The presence of N-linked glycosylation sites on the sequences was checked by using the NetNGly 1.0, a server for N-glycosylation prediction (https://services.healthtech.dtu.dk/service.php?NetNGlyc-1.0).

Database formulation and architecture

ArVirInd has a relational database structure setup using MySQL which facilitates storage, query, and visualization of information. Hyper Text Markup Language (HTML) and Cascading Style Sheet (CSS) have been used for the web-interface design. Annotated sequences data is stored in tables. The application layer between the backend tables and the web interface has been implemented by using PHP. Figure 1 explains the overall architecture of the ArVirInd knowledgebase.

Dashboard

The ArVirInd dashboard was designed using Google Data Studio (https://datastudio.google.com/) with the connectivity enabled by the ArVirInd MySQL.
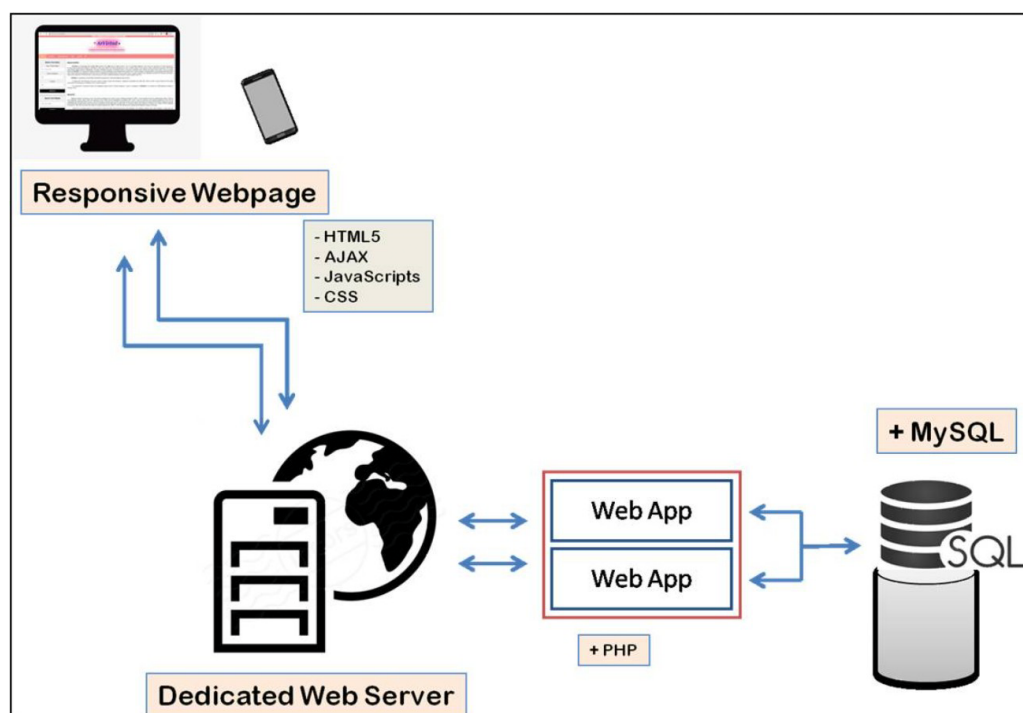
Figure 1 Data processing pipeline for ArVirInd.

Full-size ▣ DOI: 10.7717/peerj.13851/fig-1

RESULTS

Database availability and search

The ArVirInd database is open access, annotated, and curated collection of publicly available arboviral antigenic proteins from India and its neighboring countries in the subcontinent. The ArVirInd knowledgebase is available at: http://www.arvirind.co.in.

ArVirInd homepage provides the user the following ways to search: (a) search by "Virus Name", (b) search by "Country of Origin", (c) search by "Year of Origin" or a combination of these options. This is a unique feature of the database. A query returns a tabular output which gives the list of all entries in the searched category. From this table user can choose any particular entry by clicking on the respective "ArVirInd ID". The record for each entry can be printed or downloaded in PDF format. The display of the ArVirInd homepage is shown in Fig. 2A.

ArVirInd also allows the user to search (Fig. 2B) the database by submitting a query peptide (or epitope) through the "Search Your Peptide" search box (Fig. 2C). AJAX scripting is implemented to search option, which facilitates auto-search and auto-suggest query for the user.

The database features

Currently, ArVirInd consists of >400 records of antigenic proteins of Dengue, Chikungunya, Japanese encephalitis, West Nile, and Chandipura viruses from the countries:

Figure 2 ArVirInd user interfaces: (A) homepage and search, (B) display of search record, (C) result of an epitope search as returned by the 'Search Your Peptide' option, (D) database record entry. To search type keywords in relevant fields in the 'Search your Query'. This returns the list of records available. Clicking on the ArVirInd ID (blue font) will lead to the database record. User can paste the query peptide sequence to search for epitopes using the 'Search Your Peptide' option.

Full-size ☑ DOI: 10.7717/peerj.13851/fig-2

India, Bangladesh, Bhutan, Sri Lanka, Maldives, Nepal, and Pakistan. Users can also search by peptide for B-cell epitopes.

ArVirInd is a unique resource that provides annotated antigenic proteins from arboviruses isolated from the countries of the Indian subcontinent. To start with we concentrated on viral proteins that can generate B-cell and T-cell responses. For example, Dengue virus membrane glycoprotein (E-protein), and non-structural protein 1 (NS1) are known to generate antibodies (*Jayathilaka et al., 2018*). Record for each entry (sequence) provides information such as unique sequence identifier, accession number, virus and protein name, protein description, source organism, host, year of collection (outbreak), country of origin, the amino acid sequence in FASTA format, the list of B-cell epitopes, antigenic potential (VaxiJen score) (*Doytchinova & Flower, 2007*), N-glycosylation sites and database cross-references. The details of information fields in the ArVirInd entry record have been provided in the Supplementary Materials. Figure 2D shows the records of an entry in ArVirInd. This database enlists entries by the date (year) of sample collection or outbreak.

At the time of manuscript submission, we had covered strains isolated over 66 years from 1955 to 2021. ArVirInd, being a curated database, relied on complete amino acid

sequences as available in the NCBI. To the best of our knowledge, no arboviral antigenic protein sequences were available in the NCBI for India, Nepal, Bhutan, Bangladesh, Sri Lanka, Maldives, and Pakistan for the year 2022 (as the date of sample collection).

Submission of data to ArVirInd

The ArVirInd database provides an opportunity for the user to submit sequences of viral strains isolated from emerging outbreaks of Dengue, Chikungunya, West Nile, Japanese encephalitis, and Chandipura. The sequence submission tool is enlisted on the submission page. The user has to register with a valid institutional email id. This will be verified by our team before confirmation of registration. Then the user can login and submit the sequences using our sequence submission tool called, SeqKosh. All user-submitted sequences will be processed, analyzed, annotated, and uploaded by the ArVirInd team. Figure 3 indicates the sequence submission pages.

Analyses tools

The ArVirInd database has an epitope mapping tool called "EpiMapAr", which helps display the predicted and known B-cell epitopes (antigenic regions) on the database sequences. This tool can be accessed on the "EpiMapAr" page. This gives the user the choice of "virus" and "antigenic protein". As an output, it displays all the available sequences in the database for the selected category (Fig. 4).

Database dashboard

An interactive dashboard for the ArVirInd database is also available at http://arvirind.co.in/dashboard/. These summaries of the country-wise, virus-wise, and protein-wise records count information about the database (Fig. 5).

Additional information

The ArVirInd database also provides information about the list of publications of ICMR-NIV, Pune on Dengue, Chikungunya, Japanese Encephalitis, West Nile viruses, and vectors from 1957 to 2020. The 'important links' page provides links to other databases and websites important for health sciences and arboviruses. The Help page, describes basic operations for searching, viewing, downloading data, how to use the tools, dashboard, and protocol for sequence submission. The database has been tested for various browser compatibity and was found effiecient for all browsers. We recommend using the Google Chrome browser.

DISCUSSION

ArVirInd is a first-of-its-kind knowledge base dedicated to antigenic proteins from arboviral strains isolated from India and neighboring countries (the Indian subcontinent). This database enlists sequences by year of collection/isolation which coincided with the year of the outbreak. ArVirInd knowledgebase allows users to search sequences by the year of collection/isolation and country of origin, which is a unique feature. We have carried out exhaustive literature reviews to trace the year of origin for each strain. For each sequence entry, we have carried out *in silico* antigenic characterization using bioinformatics tools.

Figure 3 Sequence submission to ArVirInd: (A) User registration; (B) SeqKosh.

The summary of the findings is available in the database records for each entry. The user can also search by peptide against a record of 8,398 epitopes covering more than 400 amino acid sequences.

The online database also provides the user with an analyses tool. The EpiMapAr tool helps the user in understanding the epitope conservancy between strains (virus-wise) included in the database. This gives the user the choice of "virus" and "antigenic protein". As an output, it displays all the available sequences in the database for the selected category as below:

– Yellow highlighted sequence pattern represent Predicted B-cell Epitope (Kolaskar method)
– Green highlighted sequence pattern represent Confirmed Epitope (experimentally known).
– Purple highlighted sequence pattern represent Variability (sequence with mutation/s).

The dashboard provides the user with a summary of the datasets. Users can also submit sequences to the database through the "SeqKosh" sequence submission tool.

Dedicated databases exist for epitopes like the Bcipep, AbDb (*Ferdous & Martin, 2018*; *Dunbar et al., 2014*) and viruses such as the Ebola virus database (*Swetha et al., 2016*), or VIPR (*Pickett et al., 2012*). However, these databases provide specific information about epitopes (known) or pathogens, but none are the repository of strains belonging to specific geographical regions. Finding sequences by country of origin or year of outbreak/isolation

Figure 4 EpiMapAr: a tool for data analyses.

from other databases remains an uphill task. We tried to overcome this by creating a database that will enlist sequence information from arboviral strains isolated in the Indian subcontinent region. ArVirInd is user-friendly and enables the user to search for arboviral antigenic proteins by country of origin and year of outbreak/strain isolation. This is a unique feature of ArVirInd. In the future, we will expand the database by including tick-borne viruses such as Kyasanur Forest Disease and Crimean-Congo hemorrhagic fever (CCHF) viruses as well.

CONCLUSIONS

This article presents the formulation of a database of antigenic proteins from arbovirus strains from the countries of the Indian subcontinent. We anticipate that this database will be maintained and constantly expanded with information added for new and emerging strains of arboviruses. We believe that the ArVirInd knowledge base will benefit researchers working in immuno-informatics, arboviral diagnostics, and vaccinology. ArVirInd has the potential to become an important source of information on arboviral proteins for researchers working in India as well as countries like Bhutan, Nepal, Maldives, Bangladesh, Sri Lanka, and Pakistan. The database will be updated and expanded periodically.
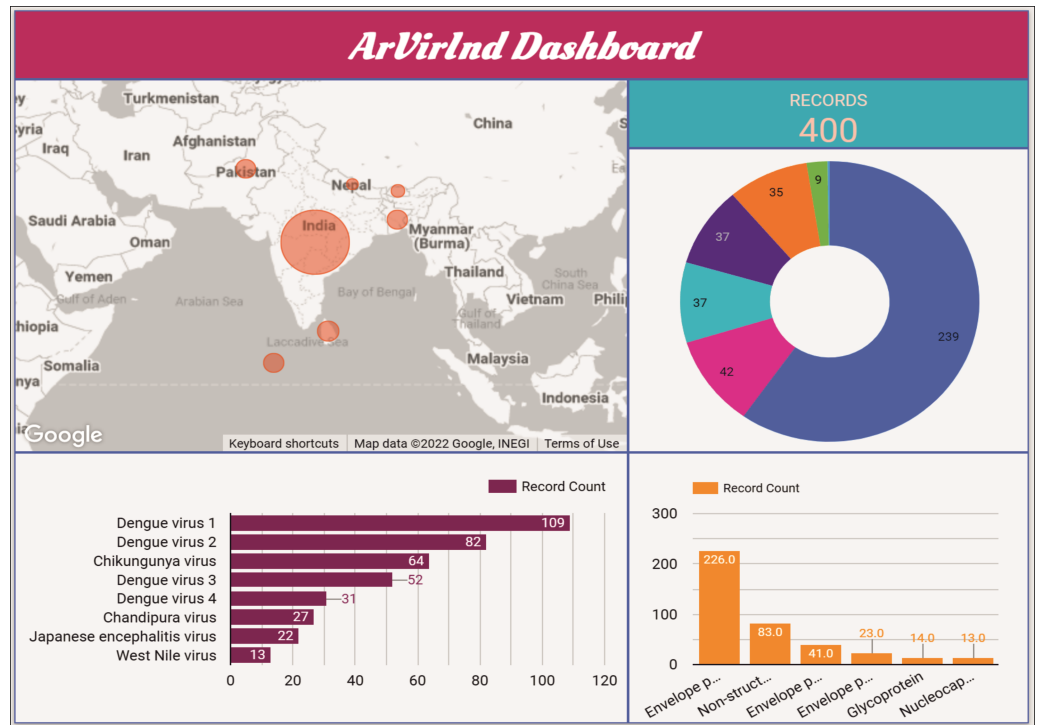
Figure 5 The ArVirInd dashboard.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Grant Disclosures

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Nitin Motilal Atre performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Kalichamy Alagarasu performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Pratip Shil conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The NCBI sequences used in the database are available in the Supplementary File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.13851#supplemental-information.

REFERENCES

Adhikari N, Subedi D. 2020. The alarming outbreaks of dengue in Nepal. *Tropical Medicine and Health* **48(1)**:5–7 DOI 10.1186/s41182-020-0194-1.

Baylis M. 2017. Potential impact of climate change on emerging vector-borne and other infections in the UK. *Environmental Health: A Global Access Science Source* **16(Suppl 1)**:112 DOI 10.1186/s12940-017-0326-1.

Dey L, Mukhopadhyay A. 2017. DenvInt: a database of protein–protein interactions between dengue virus and its hosts. *PLOS Neglected Tropical Diseases* **11(10)**:e0005879 DOI 10.1371/journal.pntd.0005879.

Dhiman RC, Pahwa S, Dhillon GPS, Dash AP. 2010. Climate change and threat of vector-borne diseases in India: are we prepared? *Parasitology Research* **106(4)**:763–773 DOI 10.1007/s00436-010-1767-4.

Doytchinova IA, Flower DR. 2007. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* **8(1)**:4 DOI 10.1186/1471-2105-8-4.

Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. 2014. SAbDab: the structural antibody database. *Nucleic Acids Research* **42(D1)**:D1140–D1146 DOI 10.1093/nar/gkt1043.

Ferdous S, Martin ACR. 2018. AbDb: antibody structure database—a database of PDB-derived antibody structures. *Database* **2018**:bay040 DOI 10.1093/database/bay040.

Gangwar RS, Shil P, Cherian SS, Gore MM. 2011. Delineation of an epitope on domain I of Japanese Encephalitis Virus envelope glycoprotein using monoclonal antibodies. *Virus Research* **158(1–2)**:179–187 DOI 10.1016/j.virusres.2011.03.030.

Gangwar RS, Shil P, Sapkal GN, Khan SA, Gore MM. 2012. Induction of virus-specific neutralizing immune response against West Nile and Japanese encephalitis viruses by chimeric peptides representing T-helper and B-cell epitopes. *Virus Research* **163(1)**:40–50 DOI 10.1016/j.virusres.2011.08.008.

Gould EA, Higgs S. 2009. Impact of climate change and other factors on emerging arbovirus diseases. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **103(2)**:109–121 DOI 10.1016/j.trstmh.2008.07.025.

Haider N, Chang Y-M, Rahman M, Zumla A, Kock RA. 2021. Dengue outbreaks in Bangladesh: historic epidemic patterns suggest earlier mosquito control intervention in the transmission season could reduce the monthly growth factor and extent of epidemics. *Current Research in Parasitology & Vector-Borne Diseases* **1**:100063 DOI 10.1016/j.crpvbd.2021.100063.

Islam Z, Mohanan P, Bilal W, Hashmi T, Rahmat Z, Abdi I, Riaz MMA, Essar MY. 2022. Dengue virus cases surge amidst COVID-19 in Pakistan: challenges, efforts and recommendations. *Infection and Drug Resistance* **15**:367–371 DOI 10.2147/IDR.S347571.

Jayathilaka D, Gomes L, Jeewandara C, Jayarathna GSB, Herath D, Perera PA, Fernando S, Wijewickrama A, Hardman CS, Ogg GS, Malavige GN. 2018. Role of NS1 antibodies in the pathogenesis of acute secondary dengue infection. *Nature Communications* **9(1)**:5242 DOI 10.1038/s41467-018-07667-z.

Mourya D, Shil P, Sapkal G, Yadav P. 2016. Zika virus: Indian perspectives. *Indian Journal of Medical Research* **143(5)**:553 DOI 10.4103/0971-5916.187103.

Nagpal BN, Gupta SK, Shamim A, Vikram K, Srivastava A, Tuli NR, Saxena R, Singh H, Singh VP, Bhagat VN, Yadav NK, Valecha N. 2016. Control of Aedes aegypti breeding: A novel intervention for prevention and control of dengue in an endemic zone of Delhi, India. *PLOS ONE* **11(12)**:1–11 DOI 10.1371/journal.pone.0166768.

Pavitrakar DV, Atre NM, Tripathy AS, Shil P. 2020. Design of a multi-epitope peptide vaccine candidate against chandipura virus: an immuno-informatics study. *Journal of Biomolecular Structure and Dynamics* **40**:1–12 DOI 10.1080/07391102.2020.1816493.

Pavitrakar DV, Damle RG, Tripathy AS, Shil P. 2018. Identification of a conserved neutralizing epitope in the G-protein of Chandipura virus. *Archives of Virology* **163(12)**:3215–3223 DOI 10.1007/s00705-018-3987-3.

Pickett B, Greer D, Zhang Y, Stewart L, Zhou L, Sun G, Gu Z, Kumar S, Zaremba S, Larsen C, Jen W, Klem E, Scheuermann R. 2012. Virus Pathogen Database and Analysis Resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. *Viruses* **4(11)**:3209–3226 DOI 10.3390/v4113209.

Reddy AS, Amarnath HSD, Bapi RS, Sastry GM, Sastry GN. 2008. Protein ligand interaction database (PLID). *Computational Biology and Chemistry* **32(5)**:387–390 DOI 10.1016/j.compbiolchem.2008.03.017.

Rijal KR, Adhikari B, Ghimire B, Dhungel B, Pyakurel UR, Shah P, Bastola A, Lekhak B, Banjara MR, Pandey BD, Parker DM, Ghimire P. 2021. Epidemiology of dengue virus infections in Nepal, 2006–2019. *Infectious Diseases of Poverty* **10(1)**:1–10 DOI 10.1186/s40249-021-00837-0.

Ryan SJ, Carlson CJ, Mordecai EA, Johnson LR. 2017. Global expansion and redistribution of Aedes-borne virus transmission risk with climate change. *BioRxiv*. 1–20 DOI 10.1101/172221.

Shil P. 2019. Rainfall and dengue occurrences in India during 2010–2016. *Biomedical Research Journal* **6(2)**:56–61 DOI 10.4103/BMRJ.BMRJ_15_19.

Shil P, Chavan SS, Cherian SS. 2011b. Antigenic variability in Neuraminidase protein of Influenza A/H3N2 vaccine strains (1968–2009). *Bioinformation* **7(1)**:76–81 DOI 10.6026/97320630007076.

Shil P, Chavan S, Cherian S. 2011a. Molecular basis of antigenic drift in Influenza A/H3N2 strains (1968–2007) in the light of antigen-antibody interactions. *Bioinformation* **6(7)**:266–270 DOI 10.6026/97320630006266.

Srinivasan E, Natarajan N, Rajasekaran R. 2020. TTRMDB: a database for structural and functional analysis on the impact of SNPs over transthyretin (TTR) using bioinformatic tools. *Computational Biology and Chemistry* **87**:107290 DOI 10.1016/j.compbiolchem.2020.107290.

Swetha RG, Ramaiah S, Anbarasu A, Sekar K. 2016. Ebolavirus database: gene and protein information resource for ebolaviruses. *Advances in Bioinformatics* **2016**:1673284 DOI 10.1155/2016/1673284.

Tabachnick WJ. 2016. Climate change and the arboviruses: lessons from the evolution of the dengue and yellow fever viruses. *Annual Review of Virology* **3(July)**:125–145 DOI 10.1146/annurev-virology-110615-035630.

Zhu M, Song X, Chen P, Wang W, Wang B. 2019. dbHDPLS: a database of human disease-related protein-ligand structures. *Computational Biology and Chemistry* **78**:353–358 DOI 10.1016/j.compbiolchem.2018.12.023.