*J Vector Borne Dis* 51, June 2014, pp. 82–85

## Research Articles

# DENVirDB: A web portal of Dengue Virus sequence information on Asian isolates

Mary J. Asnet, Amal G.P. Rubia, G. Ramya, R. Nithya Nagalakshmi & R. Shenbagarathai

*PG & Research Department of Zoology and Biotechnology, Lady Doak College, Madurai, India*

ABSTRACT

DENVirDB is a web portal that provides the sequence information and computationally curated information of dengue viral proteins. The advent of genomic technology has increased the sequences available in the public databases. In order to create relevant concise information on Dengue Virus (DENV), the genomic sequences were collected, analysed with the bioinformatics tools and presented as DENVirDB. It provides the comprehensive information of complete genome sequences of dengue virus isolates of Southeast Asia, viz. India, Bangladesh, Sri Lanka, East Timor, Philippines, Malaysia, Papua New Guinea, Brunei and China. DENVirDB also includes the structural and non-structural protein sequences of DENV. It intends to provide the integrated information on the physicochemical properties, topology, secondary structure, domain and structural properties for each protein sequences. It contains over 99 entries in complete genome sequences and 990 entries in protein sequences, respectively. Therefore, DENVirDB could serve as a user friendly database for researchers in acquiring sequences and proteomic information in one platform.

Availability: *http://www.ladydoakcollege.edu.in/denvrdb/index.php*

**Key words** Database; Dengue Virus; Genome; HMMTOP; MySQL; PHP

## INTRODUCTION

Dengue fever is one of the most important arthropod-borne viral infections that has become a major public health problem across the world. Globally, 2.5 billion people are at the risk of infection and ~5 million persons are hospitalized with dengue haemorrhagic fever (DHF) annually in Southeast Asia, the Pacific and the America[1]. Dengue epidemics have expanded in various geographical areas in the recent years. The four serotypes of dengue virus (DENV-1, DENV-2, DENV-3, and DENV-4) are identified in Asia, Africa and the America[2]. Dengue virus belongs to the genus Flavivirus within the *Flaviviridae* family. The virion particles are approximately 500Å in diam and are composed of a positive, single stranded RNA genome. The 10.8 kb genome of DENV is translated into a single long polypeptide, i.e. cleaved by proteases resulting in generation of three structural proteins—capsid (C), membrane (M) and envelope (E) glycoproteins; and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5)[3]. The capsid protein is 120 amino acids long which is involved in packaging of the viral genome and forming the nucleocapsid, whereas prM (165 aa) and envelope glycoprotein (495 aa) act as a chaperone for folding and assembly of the E protein before it is cleaved into pr peptide and M protein (~75 amino acids). The E protein contains a cellular receptor-binding site(s) for the initial binding with the host cell receptor protein[4]. Each DENV shares around 65% of the genome which is approximately the same degree of genetic relatedness as West Nile virus shares with Japanese encephalitis virus. Despite these differences, each serotype causes nearly identical syndromes in human and circulates in the same ecological niche.

Currently, there are a few databases like Flavitrack[5], and DengueInfo[6]. These databases contain genome sequences along with sequence analysis tools like pairwise and multiple sequence alignment. Moreover, the databases present only the list of sequences which are available in National Center for Biotechnology Information (NCBI) on different category. All the links are connected to NCBI database. DENVDB is yet another database developed by the National University of Singapore which provides the protein sequences of DENV in FASTA format. It would be more advantageous for the user, if the proteomic details are available in addition to the sequences in database. Hence, DENVirDB has been developed with the

objective of providing sequence information along with the computational annotation at one platform to facilitate scientists to retrieve the valuable information. It is the sequence repository of both genome and protein sequences of >90 isolates reported across nine South Asian countries.

*Database specification and acquisition*

PHP/MySQL was used to design dynamic web interface and construct a relational database to store information of complete genome sequences of dengue virus and protein sequences of each genome with annotations and results. The data consistency and non-redundancy was maintained by using normalization techniques. MySQL is capable of custom storage engines, commit grouping, gathering multiple transactions from multiple connections together to increase the number of commits per sec. The database is freely available to view and copy the available data. Genome table contains fields like Id, name, type, NCBI accession number, DENVirDB accession number, country, genome size, reference, start and end positions of each protein, protein Id and sequences in FASTA format. Protein table comprises of protein Id, country, organism, type, NCBI accession number, DENVirDB protein accession number, sequence length, theoretical pI, amino acid composition and also the secondary structure details. Protein Id acts as primary key in protein table in order to easily retrieve the data. In complete genome table, Id field is kept unique and set as primary key in order to avoid duplicate data entries and also enable to retrieve records from both the tables simultaneously. The table was arranged based on their country, type and protein name for an effortless retrieval of records. Help page is also given for the user to follow the search page.

*Database architecture*

DENVirDB provides comprehensive information on genome and protein sequences of Dengue Virus. It is created under two categories such as complete genome database for all the serotypes and protein database for all the viral proteins. The database architecture is given in Fig. 1. The primary resources of this database were retrieved from NCBI (till May 2012). The complete genome sequences of dengue virus are stored in manually designed file. The protein sequences of each nucleotide entry were computationally annotated and stored in special file format.

*Genome database*

The information of each sequence was given in the file format which contains three sections. The first part of the file presents the sequence information such as name, accession number, genome size, and organism. DENVirDB is designated with a unique accession num-

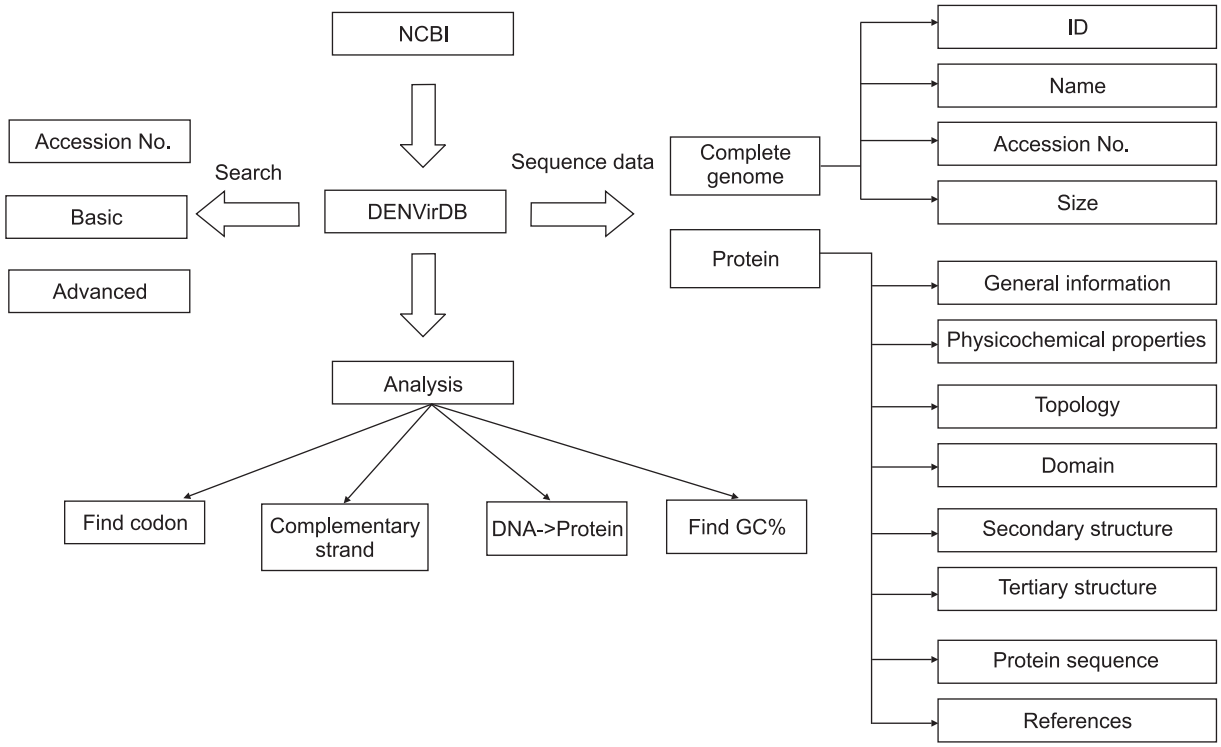
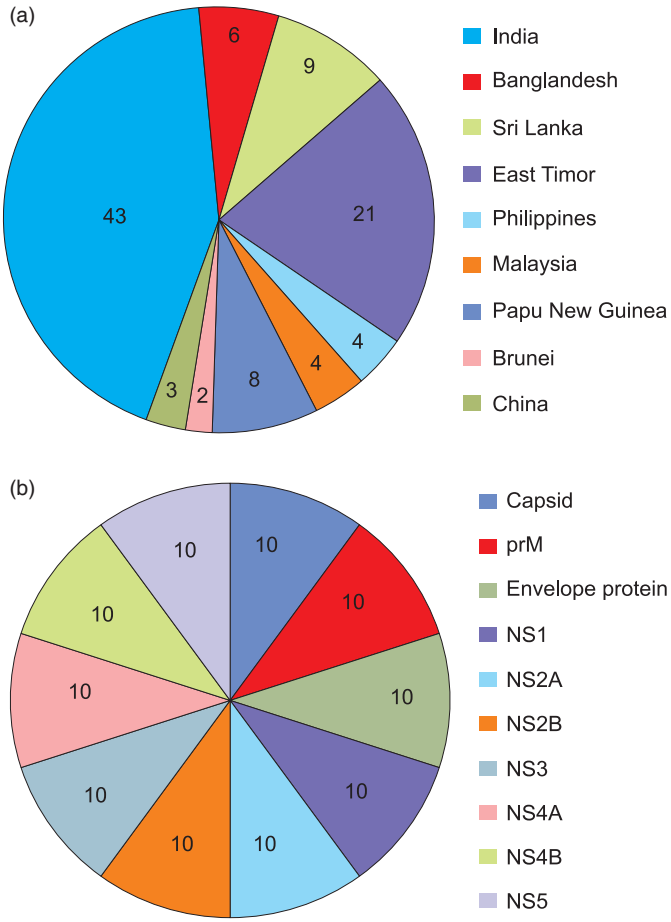



*Fig. 1:* DENVirDB database architecture.




*Figs. 2(a & b):* Database statistics: (a) DENV isolates; and (b) Structural and non-structural proteins of Dengue Virus (in percent).

ber (DEN1/2/3/4).(G/P)00000) for its easy accessibility. The second part of the file includes the sequence and genome coordinates of structural and non-structural proteins (Figs. 2a & b). The name of each protein sequence is hyperlinked to obtain its annotated results. The third part of the file contains the references. Totally, it has 98 complete genome sequences of four serotypes reported in nine countries.

*Protein database*

The complete information of computationally annotated protein sequences are given in four sections such as general information, physicochemical properties, topological features and structural details. The first section on general information encompasses the protein name, organism name, taxonomic lineage and accession number. The next section includes the physicochemical properties like molecular weight, pI, amino acid composition, and hydropathy plot. The successive section contains topological features predicted using HMMTOP[7]. The last section consists of the structural properties such as the domain details, predicted secondary structural details[8], links to three dimensional structures deposited in protein data bank (PDB), sequences available in FASTA format and references. It contains 980 proteins of both structural and non-structural protein of four serotypes of dengue virus in nine countries.

The database can be searched through basic search and advanced search. In basic genome search, nucleotide sequences are retrieved based on countries whereas in basic protein search, the selection is based on viral proteins and serotypes. Advanced genome search enables the user to search country-wise and serotype-wise. Similarly, advanced protein search provides results on the basis of viral proteins, serotypes and countries. In addition to this, the data can be accessed through the DENVirDB accession number. It also has sequence retrival option to view only the sequences in FASTA format.

*Tools*

The homepage of the database contains a few tools for sequence analysis such as FindCodon, complementary strand prediction, base composition and translation tool. In FindCodon tool, the main window consists of two text areas to obtain the nucleotide sequence and the codon for which the search is to be made. The nucleotide sequence has to be given in the text areas in complementary strand prediction and GC% prediction tool. Translation tool convert the given nucleotide sequences to mRNA sequence and then to protein sequence.

## CONCLUSION

DENVirDB is an exclusive database for dengue virus to provide genome and protein sequences, for easy accessibility of large data in a defined category. It also provides useful resource of information on the molecular and structural properties of different isolates of dengue virus in a single click. The information would also be helpful in understanding the mechanism of viral pathogenesis and the genetic diversity of different serotypes with respect to various geographical isolates. It will be periodically updated with newly deposited sequences. The database will be upgraded by including information on viral–host-vector interaction and structural motifs of viral proteins in future.

## ACKNOWLEDGEMENTS

The project was supported by Bioinformatics Infrastructure Facility (No. BT/BI/25/017/2012), Bioinformatics Division, Department of Biotechnology, Ministry

of Science and Technology, Government of India, New Delhi. The authors would like to thank Mrs N. Jeya Chithra, Head and Mrs T.R. Sivapriya, Asstt. Professor, Department of Computer Science, Lady Doak College, Madurai, for their comments and suggestions. Authors would also like to thank Mrs S. Padmaja, Asstt. Professor, Department of Physics and Mr Jason, Team Leader, A.J. Square Consultancy Services (P) Ltd. for their technical support in uploading the database in the server.

## REFERENCES

1. *Report of the Scientific Working Group on Dengue*. Geneva, Switzerland: World Health Organization 2008. WHO/TDR/SWG/08.
2. Guzman MG, Halstead SB, Artsob H, Buchy P, Farrar J, Gubler DJ, *et al.* Dengue: A continuing global threat. *Nature Rev Microbiol* 2010; *8:* S7–16.
3. Hahn YS, Galler R, Hunkapiller T, Dalrymple JM, Strauss JH, Strauss EG. Nucleotide sequence of dengue 2 RNA and comparison of the encoded proteins with those of other flaviviruses. *Virology* 1988; *162:* 167–80.
4. Perera R, Kuhn RJ. Structural proteomics of dengue virus. *Curr Opin Microbiol* 2008; *11*(4): 369–77.
5. Misra M, Schein CH. Flavitrack: An annotated database of flavivirus sequences. *Bioinformatics* 2007; *23* (19): 2645–7.
6. Schreiber MJ, Ong SH, Holland RCG, Hibberd ML, Vasudevan SG, Mitchell WP, *et al*. DengueInfo: A web portal to dengue information resources. *Infect Genet Evol* 2007; *7:* 540–1.
7. Tusnády GE, Simon I. Principles governing amino acid composition of integral membrane proteins: Applications to topology prediction. *J Mol Biol* 1998; *283:* 489–506.
8. Garnier J, Gibrat J-F, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* 1996; *266:* 540–53.

*Correspondence to:* Dr R. Shenbagarathai, Associate Professor & Head, PG & Research Department of Zoology and Biotechnology, Lady Doak College, Madurai–625 002, Tamil Nadu, India.
E-mail: ldcmadurai.btisnet@nic.in

*Received:* 29 March 2013 *Accepted in revised form:* 28 April 2014