Bioinformation by Biomedical Informatics Publishing *open access*

www.bioinformation.net **Database**

CHPVDB - a sequence annotation database for Chandipura Virus

Manas Ranjan Dikhit, Sindhu Prava Rana, Pradeep Das, Ganesh Chandra Sahoo*

BioMedical Informatics Division, Rajendra Memorial Research Institute of Medical Sciences, Agam Kuan, Patna, India - 800007; Ganesh Chandra Sahoo - E-mail: ganeshiitkgp@gmail.com; *Corresponding author

received December 10, 2008; accepted January 19, 2009; published February 28, 2009

Abstract:
Databases containing proteomic information have become indispensable for virology studies. As the gap between the amount of sequence information and functional characterization widens, increasing efforts are being directed to the development of databases. For virologist, it is therefore desirable to have a single data collection point which integrates research related data from different domains. CHPVDB is our effort to provide virologist such a one-step information center. We describe herein the creation of CHPVDB, a new database that integrates information of different proteins in to a single resource. For basic curation of protein information, the database relies on features from other selected databases, servers and published reports. This database facilitates significant relationship between molecular analysis, cleavage sites, possible protein functional families assigned to different proteins of Chandipura virus (CHPV) by SVMProt and related tools.

Availability: The database is freely available at http://chpvdb.biomedinformri.org/

Keywords: database, CHPV, CHPVDB, Chandipura virus, proteins, sequences

Background:
Chandipura virus, a member of the rhabdoviridae family and vesiculovirus genera, has recently emerged as human pathogen that is associated with a number of outbreaks in different parts of India. Although, the virus closely resembles with the prototype vesiculovirus, vesicular stomatitis virus, it could be readily distinguished by its ability to infect humans **[1]**. CHPV was noted first time in India during 1965 and was isolated from the serum of a patient with febrile illness **[2]** during an outbreak of dengue and Chikungunya viruses. Veneral transmission (which is considered as one of the modes of maintenance of the virus in nature) of Chandipura virus by males of sand fly (*Phlebotomus pappatasi* - Scopoli) has also been reported **[3]**. Chandipura virus is an enveloped RNA virus with an approximate 11,119 nucleotides nearly equal to 11 kb **[4]**. The 11 kb long genomic RNAs which code for five different structural proteins, the nucleocapsid protein (N), the phosphoprotein (P), the matrix protein (M), the glycoprotein (G) and large structural protein (L) in five monocistronic mRNAs **[5]**.

Virology is slower to embrace bioinformatics **[6]**. No computational functional analysis of different proteins of Chandipura virus is available till date. Knowledge about protein function is essential for understanding the mechanism of viral replication **[7].** Different protein functions and molecular analysis facilitates for finding potential anti-viral inhibitors. One approach for function prediction is to classify a protein into functional family. Support vector machine (SVM) is a useful method for such classification, which may involve proteins with diverse sequence distribution **[8]**. Cloning and expression of different proteins practiced by molecular biologist is helpful in restriction site analyses.

In virology research, virus-related databases and bioinformatics analysis tools are essential for discerning relationships within complex datasets about viruses **[6]**. Computational analysis on Chandipura viruses involves the general tasks related to the analysis of any novel sequences, such as molecular analysis, functional annotation, and analysis of cleavage sites of the sequences. Support vector machines (SVM), useful for predicting the functional class of distantly related proteins, is employed to ascribe a possible functional class to Chandipura virus protein **[9]**.

The large scale of protein sequences is vailable at the National Center for Biotechnology Information (NCBI) protein database and supplementary data in the published literature. The sequences of Chandipura have been downloaded from the NCBI protein database. Different strains of the same species from samples collected in different location or at different times may possess completely identical sequences. The raw dataset was preprocessed to remove the sequence smaller than 50bp while analyzing with SVMProt **[10].** The processed dataset was further refined by ProtParam **[11]** and PeptideCutter [**12**] for protein analysis.

In silico analysis give us an idea on the role of different proteins of CHPV in replication, survival and spread of CHPV in the host. Considering the biological significance of CHPV protein and with the aim of providing easy access to large and growing volume of data, we have developed CHPVDB, a repository for all known CHPV proteins. CHPVDB is a web resource, which provides sequences as well as annotation information. The CHPV protein have been analyzed, organized and integrated to develop a high user friendly database and analysis system. The web interface enables the user to execute a quick and efficient search on CHPVDB data. The database can be queried comprehensively through argument such as NCBI Locus number, different protein name, different predicted functional family and stability data. CHPVDB is an extremely useful resource for computational and experimental biologist working in related areas.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(7): 299-302 (2009)

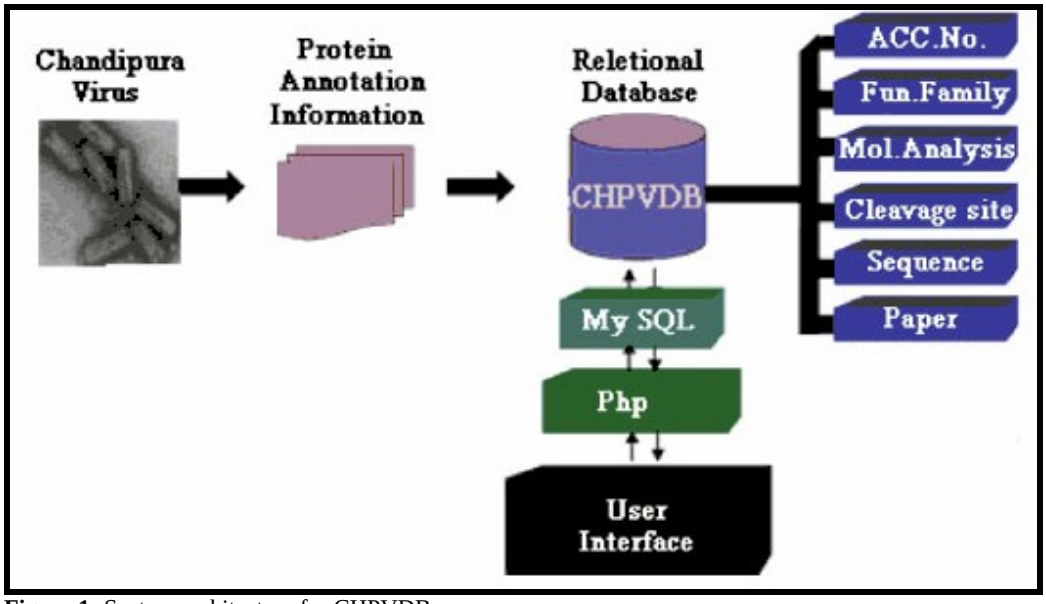

Figure 1: System architecture for CHPVDB

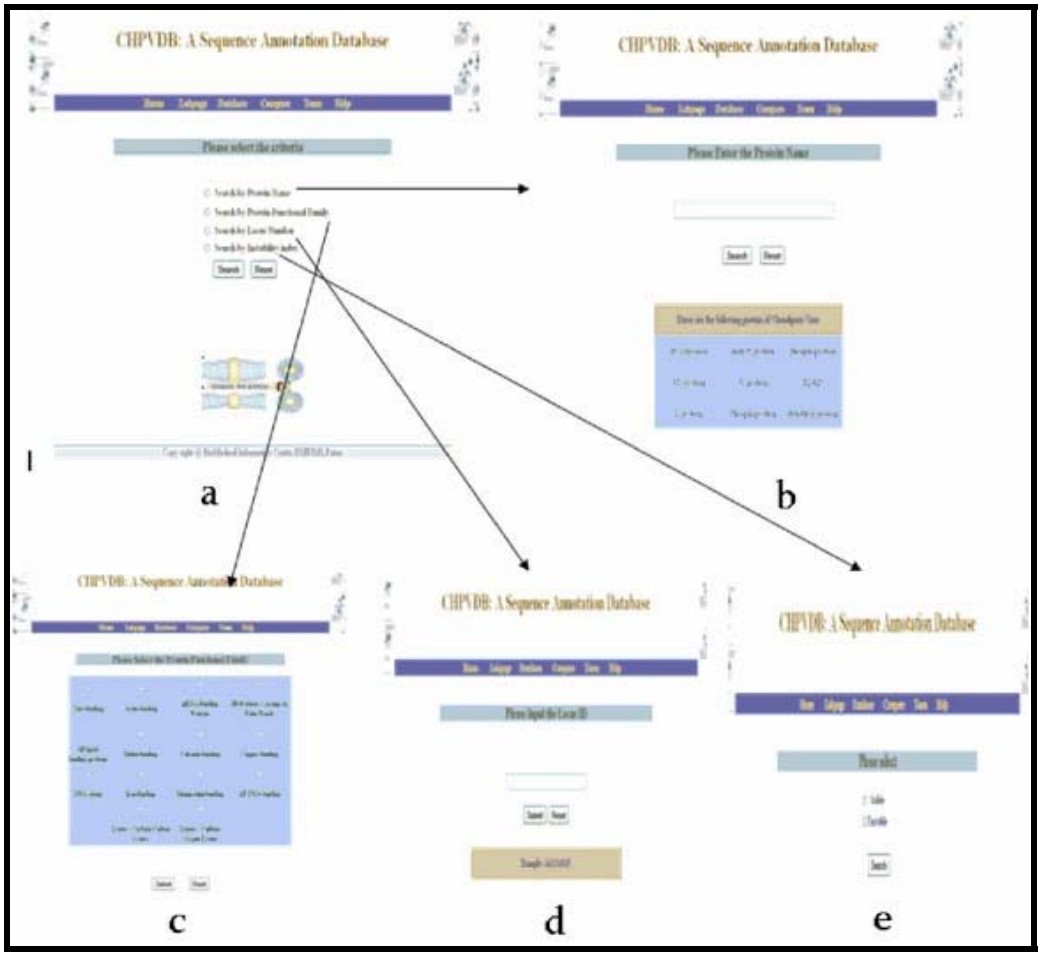

Figure 2: CHPVDB snapshot. (a) search option; (b) results by protein name; (c) results by protein family; (d) results by NCBI locus ID; (e) results by instability index.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(7): 299-302 (2009)

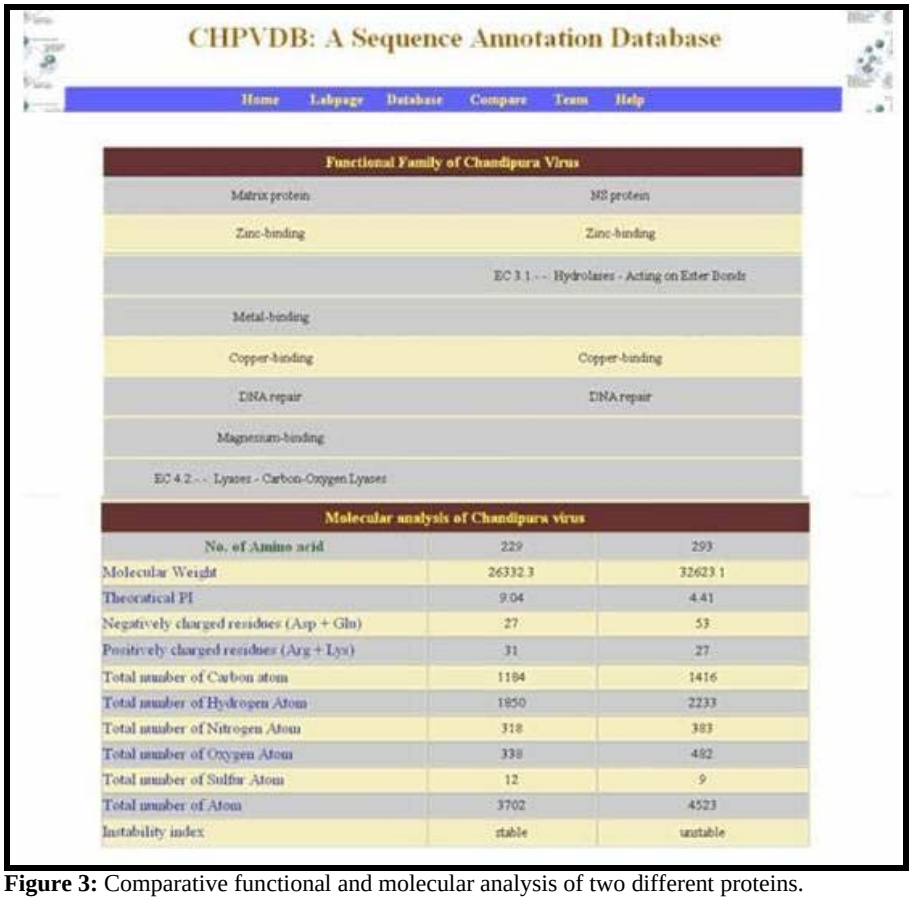

Figure 3: Comparative functional and molecular analysis of two different proteins.

Methodology:

System architecture and design

A relational database was constructed in MySQL which facilitate storage, query and visualization of annotation information. It includes three key entities: 'functional analysis', 'molecular analysis' and 'cleavage sites', for proteins. This information is managed at a protein level to provide a general view of the data. The CHPVDB data and related information are stored in MySQL relational database tables. Meta-information for different types of biological data is stored in layers of tables. The application layer between the web interface and the backend relational tables has been implemented using PHP. The overall architecture of CHPVDB is shown in **Figure 1**.

Database features

Data access

CHPVDB can be queried to obtain the information about the protein sequences in many ways. Data stored in CHPVDB can be accessed in the following ways: (i) Search by protein name: The user can enter the desired protein name to access the Meta information about the protein sequences; (ii) Search by protein functional family: The user can select the different protein functional family to find out the protein functional group of different structural and non-structural proteins; (iii) Search by NCBI locus ID: The user can enter the NCBI locus ID to obtain Chandipura virus protein sequence information; (iv) Search by Instability Index: To find out the stable and unstable protein, user can search by instability index; (v) Compare two proteins: CHPVDB can be queried to obtain the information about protein-protein comparison. The user can enter the corresponding NCBI locus ID or select the protein name to compare two proteins. Database visualization helps the user to process, interpret and act upon large stored data sets. CHPVDB provides a number of web-based forms for querying the dataset and selecting either a more detailed view of molecular annotation, cleavage 5 sites and functional family or for viewing the comparison between two selected proteins. The overall feature is shown in **Figure 2**. In an effort to improve access to diverse CHPV data, The CHPVDB has been modified to include an abundance of linkage to other database including PUBMED [**13**] for related paper abstracts and NCBI for corresponding sequences.

Data analysis

The protein function family predicted by SVMProt is different for each structural and non-structural protein of Chandipura virus strain, some of which may be responsible for virulence or pathogenicity of the virus and others for replication of the virus in the host. Prediction of the functional roles of lipid binding proteins is important for facilitating the study of various biological processes and the search for new therapeutic targets. Comparison of two amino acid sequences of any Chandipura virus protein will reveal the user, the distinguished functional properties of the corresponding protein, if there is any amino acid change at any position as SVM works on the basis of physico-chemical properties of the amino acids of the protein. In an example, when comparing functional assignment of two different proteins (RNA-dependent

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(7): 299-302 (2009)

RNA polymerase (RDRP) and matrix protein M), where functions assigned to each protein is different (functions like phosphotransferases, glycosyltransferases and mRNA capping are specific to RDRP and whereas, functions like zinc-binding, metal-binding, lyases (carbon-oxygen lyases), calcium-binding, DNA repair, copper-binding and magnesium-binding are specific for matrix protein M). It is indicated from this analysis that each protein performs a specific function assigned and evolved by the viral genome. Comparison of functions of other sequences e.g. NS protein and matrix protein reveals that zinc-binding and DNA repair functions are common to both the proteins, whereas the hydrolases (acting on ester bonds) function is specific to NS protein. However, metal-binding, magnesium-binding & carbon-oxygen lyases are specific to matrix protein (**Figure 3**). Patterns of restriction sites for all types of restriction enzymes in Chandipura virus are visualized using the web server.

Conclusion:
CHPVDB has been designed to manage and explore the vast amount of viral protein data analysis. The current version of CHPVDB has provides the information on the molecular and functional analysis of data in Chandipura virus. CHPVDB has been developed with the availability Chandipura virus proteins in public domains. We plan to include the modeled structures of different Chandipura virus proteins and analyze quantitative structure–activity relationship of novel ligands targeting different proteins in the future. The database will be updated monthly on the basis of additional data availability from analysis of the Chandipura virus sequences from other reliable resources.

Acknowledgements:
This work is supported by Indian Council of Medical Research (ICMR), Government of India. We thank Dr. Meera Singh of ICMR, New Delhi for helping us in setting up our new biomedical informatics department in RMRIMS, Patna, India. The authors would like to thank Mr. Priya Darsan Sahu for helpful discussion and valuable suggestions.

References:
[1] S. Basak *et al., Biosci. Rep.,* 27:275 (2007) [PMID: 17610154]
[2] P. N. Bhatt & F.M. Rodrigues, *Indian J. Med. Res.,* 55:1295 (1967). [PMID: 4970067]
[3] M. S. Mavale *et al., Am. J. Trop. Med. Hyg.,* 75: 1151 (2006) [PMID: 17172384]
[4] C. Marriott, *Arch. Virol.,* 150:671 (2005) [PMID: 15614433]
[5] V. A. Arankalle *et al., Emerg. Infect. Dis.,* 11:123 (2005)[PMID: 15705335]
[6] Q. Yan, *In Silico Biol.,* 8(2):71 (2008) [PMID: 18928197]
[7] D. Eisenberg *et al., Nature* 405: 823 (2000) [PMID: 10866208]
[8] Z. Cai *et al., Nucleic Acids Res.* 31: 3692 (2003) [PMID: 12824396]
[9] G. C. Sahoo *et al., Bioinformation* 3: 1 (2008). [PMID: 19052658]
[10] http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi
[11] http://www.expasy.org/tools/protparam.html
[12] http://www.expasy.org/tools/peptidecutter/
[13] www.ncbi.nlm.nih.gov

Edited by P. Kangueane
Citation: Dikhit *et al.,* Bioinformation 3(7): 299-302 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(7): 299-302 (2009)