G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics Informm 2013;11(4):289-291
http://dx.doi.org/10.5808/GI.2013.11.4.289

APPLICATION NOTE

hpvPDB: An Online Proteome Reserve for Human Papillomavirus

Satish Kumar[1]*, Lingaraja Jena[1], Sangeeta Daf[2], Kanchan Mohod[1], Peyush Goyal[3], Ashok K. Varma[4]

[1]Bioinformatics Centre and Biochemistry, Mahatma Gandhi Institute of Medical Sciences, Sevagram 442-102, India, [2]Datta Meghe Institute of Medical Sciences (Deemed University), Nagpur 440-022, India, [3]Department of Biotechnology, Ministry of Science and Technology, New Delhi 110-003, India, [4]Advanced Centre for Treatment, Research and Education in Cancer, Khargar 410-210, India

Human papillomavirus (HPV) infection is the leading cause of cancer mortality among women worldwide. The molecular understanding of HPV proteins has significant connotation for understanding their intrusion in the host and designing novel protein vaccines and anti-viral agents, etc. Genomic, proteomic, structural, and disease-related information on HPV is available on the web; yet, with trivial annotations and more so, it is not well customized for data analysis, host-pathogen interaction, strain-disease association, drug designing, and sequence analysis, etc. We attempted to design an online reserve with comprehensive information on HPV for the end users desiring the same. The Human Papillomavirus Proteome Database (hpvPDB) domiciles proteomic and genomic information on 150 HPV strains sequenced to date. Simultaneous easy expandability and retrieval of the strain-specific data, with a provision for sequence analysis and exploration potential of predicted structures, and easy access for curation and annotation through a range of search options at one platform are a few of its important features. Affluent information in this reserve could be of help for researchers involved in structural virology, cancer research, drug discovery, and vaccine design.

Keywords: comparative modeling, DNA probes, genome, HPV, neoplasms, proteome

Availability: This online reserve is made publicly available at http://www.bicjbtdrc-mgims.in/hpvPDB/.

Introduction

Human papillomavirus (HPV), a virus from the papillomavirus family, is capable of infecting humans. About 200 different strains of HPV identified, based on DNA homology, have been found to be etiologically linked to cervical, vaginal, vulvar, penile, anal, oral, and plantar infectious lesions and cancers, as well [1, 2]. The HPV genome, a double-stranded DNA molecule, consists of 8 kilobase pairs (kbp) of nucleotides, which comprises 3 regions: 6 early open reading frames (ORFs)－E1, E2, E4, E5, E6, and E7; 2 late ORFs－L1 and L2; and an upstream regulatory region [3]. A considerable volume of HPV specific information pertaining to its genome, proteome, structure, and disease association is available scattered on the web with trivial annotations; however, it is not customized to explore for data analysis, host-pathogen interaction, strain-disease association, drug designing, and sequence analysis, etc. Therefore, we proposed to develop a comprehensive reserve on HPV with maximum possible inputs and outputs for the end users.

Methods and Results

Data retrieval and curation

Amongst the existing 200 strains of HPV, 150 have been sequenced as of now, and their data available at the National Center for Biotechnology Information (NCBI). Genome and proteome information of those viral strains was retrieved from NCBI. Besides PubMed, various other online resources and published literature were also explored for validating genomic, proteomic, as well as strain and disease-associated information on HPV strains. HPV strain-specific informa-

Received September 23, 2013; Revised November 18, 2013; Accepted November 19, 2013

***Corresponding author:** Tel: +91-7152-284679, Fax: +91-7152-284038, E-mail: satishangral@gmail.com

tion, such as strain name, sequencing status, sequencing centre, NCBI accession ID, associated disease information with references, genome statistics (GC%, AT%, A, T, G, C count, genes, and proteins), etc., were curated from various online resources, and protein parameters (length, molecular weight, isoelectric point) were calculated using ExPASy ProtParam [4].

Protein structure prediction and validation

MODELLER9v10 [5] and the SWISS-MODEL [6] server were used for protein structure prediction. The stereochemistry of each protein was evaluated through PROCHECK [7] analysis, available at the RCSB validation server (http://deposit.rcsb.org/validate/), and validated using ProSA-web [8] (http://prosa.services.came.sbg.ac.at/prosa.php).

Reserve architecture and design

Human Papillomavirus Proteome Database (hpvPDB), the relational reserve, was developed using Microsoft SQL Server 2005 as the back end. The website is powered by XAMPP (Windows Version 1.7.3). HTML, JavaScript, and CGI-PERL-based web interfaces were employed to execute SQL queries. The curated data and related information were stored in tables. The application layer, the web interface, and the backend relational tables were integrated using CGI-PERL. The overall architecture of hpvPDB is shown in Fig. 1.

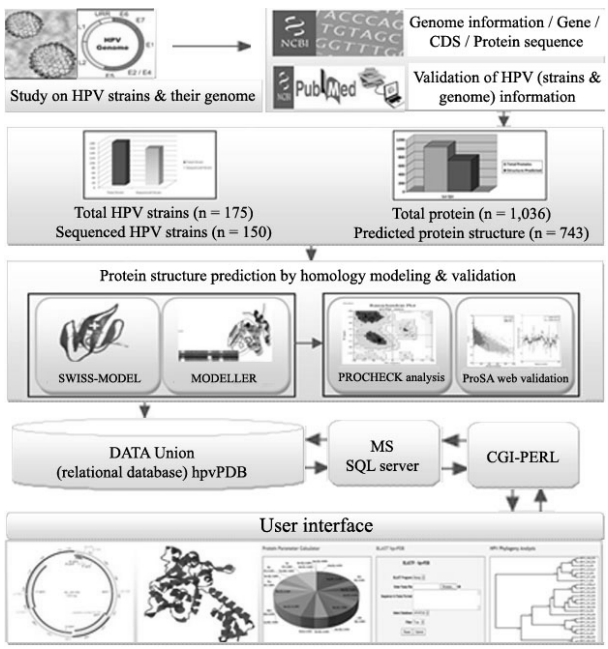

Fig. 1. System architecture of Human Papillomavirus Proteome Database (hpvPDB) showing data collection, analysis, union, and user interface. HPV, human papillomavirus; CDS, coding sequence.

Reserve features

hpvPDB interfaces are made to help the users for easy navigation and information retrieval. Home, About, Tools, Search, and Advanced Search interfaces can be explored to obtain strain- and protein-specific information. User can access the meta information about different strains using a search box. Reserve comprises the strain-specific detailed informationon its name, sequencing status, submission details, date of submission, NCBI IDs, disease types and subtypes, type of DNA, genome length, molecular weight, nucleotide composition (A, T, G, C, AT, GC content), number of genes and proteins, and protein list. A genome map of each strain obtained by Geneious 5.4.4 software (available from http://www.geneious.com/) is also integrated in this page. Users, through an advanced search option, can precisely access the Genome and Proteome information separately by selecting HPV genome or HPV proteome. Each protein entry comprises protein overview (name, locus, function, etc.), protein sequence information (amino acid sequences with NCBI accession number with provision for direct protein BLAST [9] against NCBI nr database), protein parameters (length, molecular weight, theoretical isoelectric point [pI], amino acid composition, etc.), protein structure (predicted 3D structure by homology modeling viewed by Jmol (available from http://www.jmol.org/) [10] with the JAVA platform, Ramachandran plot obtained by PROCHECK and Z-score and Energy plot obtained by ProSA-Web. hpvPDB platform also provides a phylogeny analysis tool to perform multiple sequence alignment and phylogenetic tree construction of selected HPV proteins using the Phylogeny.fr web service [11].

The original Human Papillomaviruses Database was developed and hosted by the Los Alamos National Laboratory (LANL) between 1994 and 1999 with funding from the National Institute of Allergy and Infectious Diseases (NIAID) [12]. 'Human Papillomaviruses: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences' contains four annual data books of papillomavirus information published in both paper and electronic form (1994, 1995, 1996, and 1997) but has not been updated since 1997 [12]. This contains nucleotide sequences of few HPV strains and other papillomaviruses, amino acid and nucleotide sequence alignments, analysis, related host sequences, and database communication. We did not find any structural information in that database. In hpvPDB, along with updated protein sequence information, genome and protein structure information is also provided.

Conclusion

hpvPDB brings together comprehensive information on a total of 1,036 protein sequences and 743 predicted structures. The outcome of this study might provide a platform for simultaneous structural comparative analysis of these proteins and help in finding out variations in their structures to explore why different strains of HPV have causative associations with different types of cancers. Further, this might also help in designing specific drugs or vaccines against specific strains of HPV. This reserve provides a resource to help virologists identify potential roles for viral protein. Currently the hpvPDB is updated manually through online resources and available scientific publication review; however, to sustain the quality, these data are analyzed and checked before incorporation into this reserve. Meanwhile, to provide regular updates, our team is committed to searching for newly sequenced HPV strains, updating this reserve, and serving the users.

Acknowledgments

Authors express gratitude to the Department of Biotechnology, MoS&T, Government of India for their financial support to Bioinformatics Centre, wherein this online reserve has been developed. Authors thank Dr. B.C. Harinath, Director, JBTDRC and Coordinator, Bioinformatics Centre for his insightful comments and suggestions. Grateful thanks to Shri D.S. Mehta, President, Kasturba Health Society; Dr. (Mrs.) P. Narang, Secretary, Kasturba Health Society; Dr. B.S. Garg, Dean, MGIMS; and Dr. S.P. Kalantri, MS, Kasturba Hospital, MGIMS, Sevagram for their encouragement and unconditional support.

References

1. Tungteakkhun SS, Filippova M, Neidigh JW, Fodor N, Duerksen-Hughes PJ. The interaction between human papillomavirus type 16 and FADD is mediated by a novel E6 binding domain. *J Virol* 2008;82:9600-9614.
2. Watts KJ, Thompson CH, Cossart YE, Rose BR. Variable oncogene promoter activity of human papillomavirus type 16 cervical cancer isolates from Australia. *J Clin Microbiol* 2001;39: 2009-2014.
3. Zheng ZM, Baker CC. Papillomavirus genome structure, expression, and post-transcriptional regulation. *Front Biosci* 2006;11:2286-2302.
4. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, *et al*. *The Proteomics Protocols Handbook*. Totowa: Humana Press, 2005. pp. 571-607.
5. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, *et al*. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci* 2007; Chapter 2:Unit 2.9.
6. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 2006;22:195-201.
7. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 1993;26:283-291.
8. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* 2007;35:W407-W410.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403-410.
10. Jmol: an open-source Java viewer for chemical structures in 3D with features for chemicals, crystals, materials and biomolecules. Jmol. Accessed 2013 Feb 5. Available from: http://www.jmol.org/.
11. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, *et al*. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 2008;36:W465-W469.
12. Myers G, Baker C, Wheeler C, Halpern A, McBride A, Doorbar J. Human papillomaviruses: a compilation and analysis of nucleic acid and amino acid sequences. Los Alamos: Theoretical Biology and Biophysics, Los Alamos National Laboratory, 1994-1997. Accessed 2013 Nov 10. Available from: http://pave.niaid.nih.gov/lanl-archives/HPVcompintro4. html.