Nucleic Acids Research, 2004, Vol. 32, Database issue ***D289–D292***
DOI: 10.1093/nar/gkh098

VirGen: a comprehensive viral genome resource

Urmila Kulkarni-Kale*, Shriram Bhosle, G. Sunitha Manjari and A. S. Kolaskar[1]

Bioinformatics Centre and [1]Vice-Chancellor's office, University of Pune, Pune 411 007, India

Received August 14, 2003; Revised September 21, 2003; Accepted October 8, 2003

ABSTRACT

VirGen is a comprehensive viral genome resource that organizes the 'sequence space' of viral genomes in a structured fashion. It has been developed with the objective of serving as an annotated and curated database comprising complete genome sequences of viruses, value-added derived data and data mining tools. The current release (v1.1) contains 559 complete genomes in addition to 287 putative genomes of viruses belonging to eight viral families for which the host range includes animals and plants. Viral genomes in VirGen are annotated using sequence-based Bioinformatics approaches. The genomic data is also curated to identify 'alternate names' of viral proteins, where available. VirGen archives the results of comparisons of genomes, proteomes and individual proteins within and between viral species. It is the first resource to provide phylogenetic trees of viral species computed using whole-genome sequence data. The module of predicted B-cell antigenic determinants in VirGen is an attempt to link the genome to its vaccinome. Comparative genome analysis data facilitate the study of genome organization and evolution of viruses, which would have implications in applied research to identify candidates for the design of vaccines and antiviral drugs. VirGen is a relational database and is available at http://bioinfo.ernet.in/virgen/virgen.html.

INTRODUCTION

Genome sequencing projects of various organisms have transformed biology into a data-rich information science. Viruses are the group of organisms that have the largest numbers of genomes sequenced and these sequences are deposited in the public domain sequence databases. Although there exist numerous genomic databases/resources for microbial and model organisms (1), such resources for viruses are unavailable.

A few virus-specific databases such as VIDA (2), VGDB (3) and the Sub Viral RNA database (4) have been developed and are available online. VIDA archives the open reading frames of animal viruses and VGDB stores genes and proteins of about 21 pox and related viruses. The Sub Viral RNA database contains partial and complete sequence data of the smallest known auto-replicable species. The viral genome resources at the NCBI and at the EBI list only ~1600 and ~900 records, respectively (5,6) even though the primary repositories of nucleic acid sequences contain a larger number of entries for the same. Viral genome data, if organized in a structured fashion, would facilitate navigation through the large 'sequence space' and offer several opportunities for data mining and knowledge discovery.

WHAT IS VirGen?

VirGen comprises primary and derived data of viral genomes and is designed to serve as a single-stop solution for accession, retrieval and analysis of viral genome sequences. Unique features of VirGen include easy access to curated and annotated viral genome records, graphical representation of genome organization, a set of non-redundant genome records, precomputed data on multiple alignments of genomes/proteomes and predicted antigenic determinants. VirGen is the first database to curate viral genomes for 'alternate names' of proteins and to archive the results of whole-genome phylogeny. Primary as well as derived data in VirGen are organized as a relational database with a user-friendly front-end and software tools for data analysis and mining.

DATABASE DESCRIPTION

Design and implementation

VirGen consists of an administrative module for data acquisition and curation along with three integrated sections for genome annotation, genome analysis and prediction of antigenic determinants, as shown in Figure 1. The data are organized in the relational database management system MySQL, on a Microsoft Windows 2000 server. The query system has been developed using CGI Perl scripts and ASP. A user-friendly web interface was designed in HTML by implementing VB and Java scripts. Parsing, annotation and data updates have been automated to minimize human intervention.

Genome data acquisition

The viral genome sequences are obtained from the repository of nucleic acid sequences available at the NCBI server (5). An automated tool to retrieve the genomic sequences of viruses has been developed that uses keywords and MeSH terms such as genome, genomic, etc. The genome data are further filtered

*To whom correspondence should be addressed. Tel: +91 20 569 0195/ 569 2039; Fax: +91 20 569 0087; Email: urmila@bioinfo.ernet.in
Correspondence may also be addressed to A. S. Kolaskar. Email: kolaskar@bioinfo.ernet.in

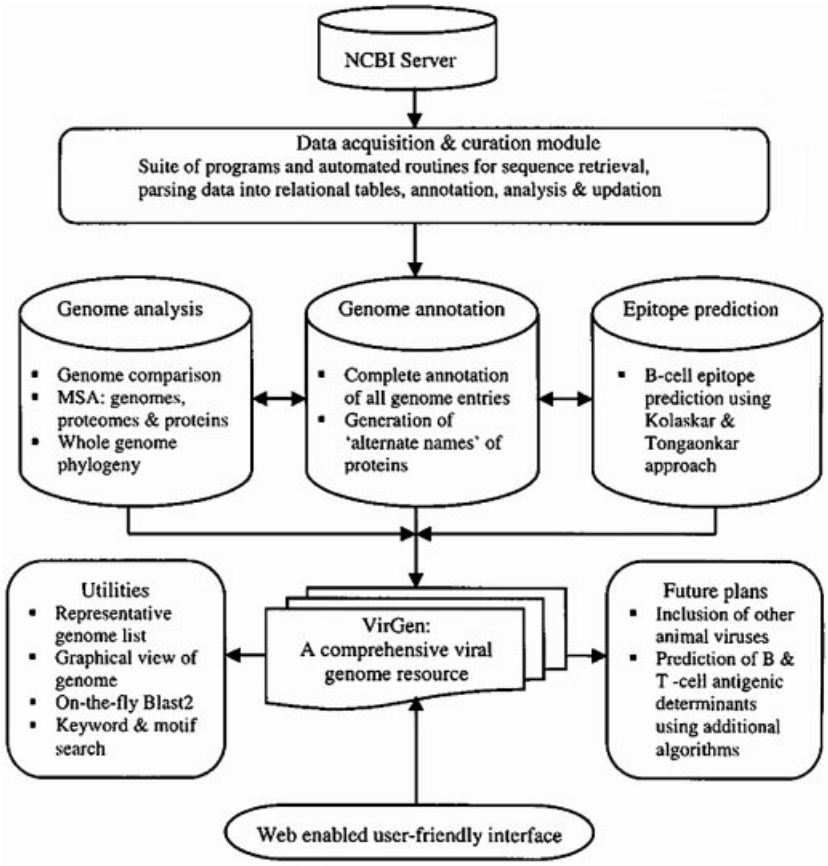

Figure 1. Organization of VirGen.

based on database division, type of genetic material, genome size, etc. The known complete genome entries and entries that are not annotated as 'complete genome records' but are likely to be complete genomes are also downloaded, as the sequence length of such entries is in the range typical of complete genome records for the respective viruses. Such genomic entries are made available in VirGen as 'putative genomes'. Version 1.1 of VirGen archives the data of eight viral families with positive sense ssRNA genome.

Processing of genomic data

The genome entries are organized hierarchically from family, subfamily, genus, species to strain/isolate level (7,8). A well-annotated genome record of a characterized strain/isolate of each viral species is identified as a 'representative entry'. The set of representative entries and abbreviations of virus names in VirGen are consistent with those listed by the International Committee on Taxonomy of Viruses (ICTV) (7). Representative entries provide a non-redundant set of viral genome sequences, which are subsequently used for annotation and to study phylogenetic relationships. Strain/isolate information was added manually by browsing the cited references, wherever possible. The genome data is also curated with respect to the nomenclature and latest taxonomic status.

Genome annotation

Every entry in the VirGen database is annotated with reference to its genome organization. The entries are annotated in terms of individual coding sequences (cds), polyprotein(s) and individual proteins using representative genomes and the program BLAST v2.2.5 (9). For example, as shown in Figure 2, the SARS coronavirus codes for two polyproteins, namely ORF1ab and ORF1a. If annotations are not available for the representative entry, then the protein sequences of the virus from the PIR-PSD (10) and/or Swiss-Prot (11) database were used to annotate its genomic record using TBLASTN. The cut-off values used were: sequence identity >95% and length of aligned sequence equal to the length of the protein sequence.

Generating the list of alternative names

During the process of annotation, it was observed that the same protein was referred to by various names among viral species belonging to the same genus. Such nomenclature problems in biology are well known and attempts are being made to standardize the vocabulary of terms at various levels of biocomplexity (12). We have used protein sequence as a probe to identify 'alternate names', which are displayed along with the annotations (see Fig. 2). These alternate names are treated as synonyms while processing keyword-based searches in VirGen. The dictionary of alternate names could also be used to compile a controlled vocabulary for viral proteins.

Schematic representation of genome organization

The annotated entries are used to dynamically generate and display the graphical view of the genome/proteome using

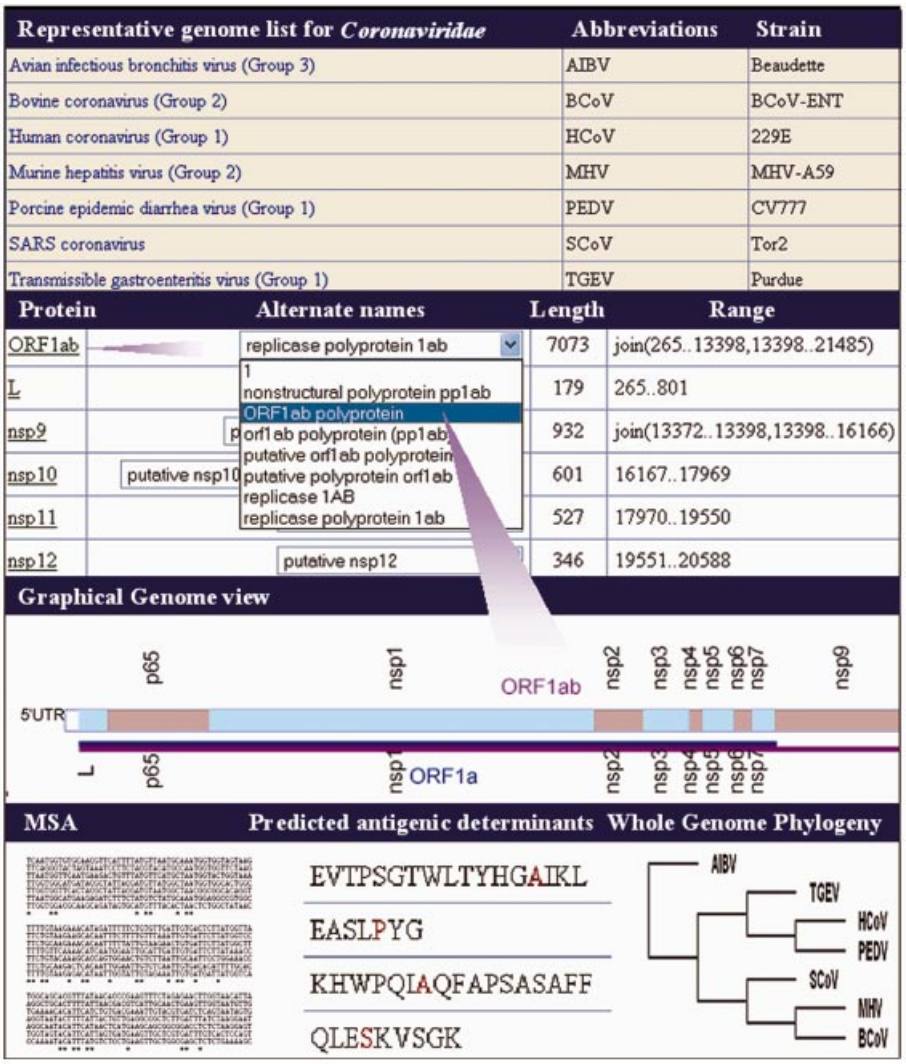

Figure 2. Composite picture of a sample session of VirGen using the SARS coronavirus as an example.

Scalable Vector Graphics (SVG). The graphical display illustrates interesting features of viral genome organization (see Fig. 2). For example, as can be seen in Figure 2, both the ORFs of the SARS coronavirus overlap significantly. The graphical view of genome organization could also be used to retrieve the sequence(s) of an entire genome/proteome or individual proteins in the FASTA format by clicking the labels.

Generation of derived data

Pairwise and multiple sequence alignment of genomes/proteomes. Divergence from a common ancestor and acquisition of new functions by variations in sequence and structure is the principle mechanism through which viruses evolve species and strain-specific properties. CGI-enabled Perl scripts have been developed to facilitate on-the-fly pairwise comparison of genomes/proteomes using the BLAST suite of programs (9). Multiple sequence alignment (MSA) of genomes and proteomes facilitates the comparison of similarities and differences at equivalent positions in the sequences of a group of related viruses. Although a few databases provide MSAs of individual viral proteins, there are none providing MSAs of genomes and proteomes. VirGen provides MSA with high resolution and granularity by including the genomic context in non-redundant data sets (see Fig. 2). Predicted antigenic determinants when mapped on MSAs help in the identification of strain- and/or species-specific antigenic regions and have implications for the design of viral vaccines (13,14). The parallel version of ClustalW v1.8 [Thompson and coworkers and Silicon Graphics Inc. (SGI)], implemented on an SGI Onyx 300, a four-processor parallel machine, has been used to obtain MSA data (15).

Whole-genome phylogeny. Whole-genome phylogenetic studies are expected to represent the true phylogenetic relationship observed between closely related viruses as compared with studies using a single gene or protein. Phylogenetic studies carried out using complete genome data of viruses are relevant because the overall genome organization of viruses is found to be conserved among members of the same family. Whole-genome phylogeny, though preferred, is seldom used due to its compute-intensive

nature. VirGen caters to this need by computing whole-genome phylogenetic trees using a parallel version of PHYLIP v3.573 (J. Felsenstein, Department of Genetics, University of Washington, Seattle and SGI) implemented on an SGI Onyx 300. Unrooted phylogenetic trees are obtained using both DNA and protein parsimony methods (16). The statistical significance of these trees is evaluated by bootstrapping the sequences to generate *n* datasets, such that *n* is equal to one-sixth of the sequence length. A consensus tree is derived for representative entries from the bootstrap data of *n* phylogenetic trees (see Fig. 2). Family- and genus-specific unrooted phylogenetic trees are part of VirGen. Species-specific phylogenetic analyses enable the study of the evolution of various strains belonging to a species and these trees will be included in future releases of VirGen.

Prediction of antigenic determinants. VirGen archives predicted B-cell antigenic determinants using the method of Kolaskar and Tongaonkar (17). The predictions are restricted to known antigenic proteins (7). Algorithms used for the prediction of B-cell antigenic determinants help to reduce the solution space to a few peptides as compared with the conventional methods that involve sequencing of overlapping peptides to determine the antigenic regions. This compilation provides easy and timely access to predicted antigenic determinants.

HOW TO BROWSE VirGen

Logical navigation through the family up to strain/isolate level is provided in addition to keyword-based and sequence-based search utilities. Genome records are displayed alphabetically within the genus. The data sets of MSA, phylogeny and predicted antigenic determinants can be browsed independently. Documentation, online help and a guided tour are included. The guided tour provided in the Help illustrates various features/utilities of VirGen using the SARS coronavirus as an example.

AVAILABILITY OF VirGen

The current release (v1.1, September 20, 2003) archives 559 complete genomes in addition to 287 putative genomes of viruses that belong to families *Arteriviridae*, *Astroviridae*, *Coronaviridae*, *Dicistroviridae*, *Flaviviridae*, *Luteoviridae*, *Picornaviridae* and *Togaviridae*. A bimonthly updating schedule is planned. VirGen has been developed at the Bioinformatics Centre, University of Pune and can be accessed through http://bioinfo.ernet.in/virgen/virgen.html.

FUTURE PLANS

Complete genome data of various viral families and additional methods for prediction of B- and T-cell antigenic determinants will be included in future releases of VirGen. Species-specific phylogenetic trees will also be computed and bootstrap data of all unrooted phylogenetic analyses will be made available through an anonymous ftp.

ACKNOWLEDGEMENTS

The contributions of Mr Avinash Gill and Dr Pushparaj Madavi are acknowledged. VirGen is funded by the Department of Biotechnology, Government of India.

REFERENCES

1. Baxevanis,A.D. (2003) The molecular biology database collection: 2003 update. *Nucleic Acids Res.*, **31**, 1–12.
2. Alba,M.M., Lee,D., Pearl,F.M.G., Shepherd,A.J., Martin,N., Orengo,C.A. and Kellam,P. (2001) VIDA: a virus database system for the organization of animal virus genome open reading frames. *Nucleic Acids Res.*, **29**, 133–136.
3. Hiscock,D. and Upton,C. (2000) Viral genome database: storing and analyzing genes and proteins from complete viral genomes. *Bioinformatics*, **16**, 484–485.
4. Pelchat,M. (2003) SubViral RNA: a database of smallest known auto-replicable RNA species. *Nucleic Acids Res.*, **31**, 444–445.
5. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler, G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
6. Brooksbank,C., Camon, E., Harris,M.A., Magrane,M., Martin,M.J., Mulder,N., O'donovan,C., Parkinson,H., Tuli,M.A., Apweiler,R. *et al.* (2003) The European Bioinformatics Institute's data resources. *Nucleic Acids Res.*, **31**, 43–50.
7. Van Regenmortel,M,H.V., Fauquet,C.M., Bishop,D.H.L., Carstens,E.B., Estes,M.K., Lemon,S.M., Maniloff,J., Mayo,M.A., McGeoch,D.J., Pringle,C.R. *et al.* (Eds) (2000) *Virus Taxonomy: Classification and Nomenclature of Viruses. Seventh Report of the International Committee on Taxonomy of Viruses*. Academic Press, San Diego, CA.
8. Büchen-Osmond,C. (2003) The Universal Virus Database ICTVdB. *Comput. Sci. Eng.*, **5**, 16–25.
9. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Wu,C.H., Yeh,L.-S.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
11. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher, A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight, S.S., Eppig,J.T. *et al.* The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
13. Kolaskar,A.S. and Kulkarni-Kale,U. (1999) Prediction of three-dimensional structure and mapping of conformational antigenic determinantss of envelope glycoprotein of Japanese encephalitis virus. *Virology*, **261**, 31–42.
14. Kulkarni-Kale,U. and Kolaskar,A.S. (2003) Prediction of 3D structure of envelope glycoprotein of Sri Lanka strain of Japanese encephalitis virus. In Yi-Ping Phoebe Chen (ed.), *Conferences in Research and Practice in Information Technology*, Australian Computer Society Inc., Sydney, Australia, vol. 19, pp. 87–96.
15. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
16. Felsenstein,J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.*, **266**, 418–427.
17. Kolaskar,A.S. and Tongaonkar,P.C. (1990) A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett.*, **276**, 172–174.