D382–D385 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj023

DPVweb: a comprehensive database of plant and fungal virus genes and genomes

Michael J. Adams* and John F. Antoniw

Plant-Pathogen Interactions Division, Wheat Pathogenesis Programme, Rothamsted Research, Harpenden, Herts AL5 2JQ, UK

Received August 4, 2005; Revised and Accepted September 21, 2005

ABSTRACT

DPVweb (http://www.dpvweb.net/) provides a central source of information about viruses, viroids and satellites of plants, fungi and protozoa. Comprehensive taxonomic information, including brief descriptions of each family and genus, and classified lists of virus sequences are provided. The database also holds detailed, curated, information for all sequences of viruses, viroids and satellites of plants, fungi and protozoa that are complete or that contain at least one complete gene (currently, $n \approx 9000$). For comparative purposes, it also contains a single representative sequence of all other fully sequenced virus species with an RNA or single-stranded DNA genome. The start and end positions of each feature (gene, non-translated region and the like) have been recorded and checked for accuracy. As far as possible, nomenclature for genes and proteins are standardized within genera and families. Sequences of features (either as DNA or amino acid sequences) can be directly downloaded from the website in FASTA format. The sequence information can also be accessed via client software for PC computers (freely downloadable from the website) that enable users to make an easy selection of sequences and features of a chosen virus for further analyses.

INTRODUCTION

The public sequence databases contain vast amounts of data on virus genomes but accessing and comparing the data, except for relatively small sets of related viruses can be very time consuming. The procedure is made difficult because some of the sequences on these databases are incorrectly named, poorly annotated or redundant. The NCBI Reference Sequence project (1) provides a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA) and protein products, for major research organisms. This now includes curated information for a single sequence of each fully sequenced virus species. While this is a welcome development, it can only deal with complete sequences. DPVweb aims to provide a comprehensive source of high quality sequence and taxonomic information for all fully sequenced genes of viruses, viroids and satellites of plants, fungi and protozoa. An important feature is the opportunity to access genes (and other features) of multiple sequences quickly and accurately. Thus, for example, it is easy to obtain the nucleotide or amino acid sequences of all the available accessions of the coat protein gene of a given virus species or for a group of viruses. To increase its usefulness further, DPVweb also contains a single representative sequence of all other fully sequenced virus species with an RNA or single-stranded DNA (ssDNA) genome.

ORIGIN OF THE PROJECT

The project developed as an electronic version of the Association of Applied Biologists (AAB) descriptions of plant viruses (DPV) (2). This is a standalone program for PCs (first released in 1998) that includes detailed descriptions of selected plant viruses together with information on taxonomy and sequences. It provides a comprehensive resource that is widely used for teaching, disease management and research. As part of this program, software (DPVMap) was written to display selected virus sequences interactively. A separate enhanced feature table (EFT) file written for each sequence contains the start and end nucleotide positions of the features [e.g. open reading frames (ORFs) and untranslated regions] within the sequence. In DPVMap any of the features of the sequence can be dragged into a sequence editor to display its nucleotide sequence (as RNA or DNA), or the predicted amino acid sequence of an ORF. Annotations provide for the correct display of reverse complementary sequences and of those incorporating a frameshift or intron. Sequence features are checked for accuracy and, as far as possible, nomenclature

*To whom correspondence should be addressed. Tel: +44 1582 763133; Fax: +44 1582 760981; Email: mike.adams@bbsrc.ac.uk

for genes and proteins are standardized within genera and families to make it easier to compare features from different viruses. Additional information (e.g. polyprotein cleavage sites) may also be added by comparison with related sequences. From a modest beginning, this has expanded to become a comprehensive resource (see Current Coverage section, below). The information contained in the individual EFT files is valuable because it has been checked for accuracy and is often more detailed than that provided in the original sequence file from EMBL/Genbank/DDBJ. However, the EFT files can only be used with DPVMap and to examine one sequence at a time. We therefore decided to transfer this information together with the sequences themselves into a database table so that multiple datasets could be selected and extracted easily and then used for further analysis. Although the standalone program is still available in CD-ROM format, much of the information, including that of the sequence features, is now also provided through the website described here.

DESCRIPTION OF THE WEBSITE

The home page of DPVweb (http://www.dpvweb.net/) provides simple and user-friendly access to all the features of the site. There are five major sections to the site, accessed from a menu bar at the top of the page.

Home

This provides an overview of the site and an introduction to plant viruses in general.

DPV

This section contains the detailed descriptions of individual viruses. Over 400 viruses or virus groups are included. Numbers 1–354 were originally published in paper form by the AAB between 1970 and 1989, while additional descriptions have been added since 1998. A small editorial team of internationally renowned virologists commission and edit the descriptions, which are written by specialists. The descriptions provide comprehensive information on virus diseases, geographical distribution, host range, symptoms, transmission, vectors, serology, relationships with other viruses, purification protocols, properties of virus particles, particle structure and composition, molecular structure, genome properties, cytopathology, ecology and control procedures as well as references to the scientific literature. The descriptions also include images of symptoms, electron micrographs of virus particles, genome maps and so on. A menu item allows a search to be made using a text query on the complete set of descriptions, and this can be limited to selected fields (subheadings in the descriptions) if required.

Notes

The notes are a brief description of each family and genus included within the project (i.e. all viruses except those with a double-stranded DNA (dsDNA) genome; see Current Coverage section, below). This includes (i) a description of particle morphology and genome organization (where known); (ii) a genome map (usually of the genus type member); (iii) a representative electron micrograph (for plant viruses only); (iv) a list of species with their acronyms and synonyms for each genus with links to the plant virus descriptions; (v) a list of accession numbers (and links to the sequences) used by EMBL/Genbank/DDBJ databases for all the sequences in the family or genus together with the description from the sequence header (for viruses, satellites and viroids infecting plants, fungi and protozoa only); and (vi) a list of 'curated sequences' for each virus, providing links to access the sequences of individual genes or other features (see Access to Sequence Features section, below). A menu item allows a search to be made amongst all the virus (and strain/synonym) names in the database. This returns a list of all matching names, with links to the appropriate genus notes page and, where appropriate, to the specialist description.

Sequences

This provides lists of the accession numbers used by EMBL/Genbank/DDBJ databases for all the sequences of viruses, satellites and viroids infecting plants, fungi and protozoa. These have been checked to ensure that, as far as possible, they are allocated to their correct species. Each accession number is linked so that the sequence can be fetched from EMBL.

Analysis

This provides links to client software that accesses the data from the user's PC (see Client Software section, below).

TECHNICAL DETAILS

All data are stored in a MySQL database on an Apache web server and accessed using PHP scripts to display the data on the web page and to generate XML output for the web-enabled client software. Regular searches are made at EMBL for virus sequences that are new or that have been updated. EFT files are created for the standalone DPV program as described above (Origin of the Project section) and used to test the accuracy of the data. The information is transferred into XML database files for use by the standalone DPV program, and the public web database is then updated once a month.

CURRENT COVERAGE

The notes and taxonomic information cover all RNA and ssDNA viruses, including the reverse-transcribing viruses and related retroelements and therefore includes all viruses except those with a dsDNA genome. The sequence database holds detailed information for all sequences of viruses, viroids and satellites of plants, fungi and protozoa that are complete or that contain at least one complete gene (currently, $n \approx 9000$). For comparative purposes, it also contains a single representative sequence of all other fully sequenced virus species within the taxa covered ($\sim$750 sequences).

ACCESS TO SEQUENCE FEATURES

Curated sequences are listed under the current species name from the appropriate genus notes page. The descriptions are shorter, but usually more informative than those in the

complete sequence lists. The accession number is linked to the entire sequence file. Next to each accession number is a check box for selection of one or more sequences within the genus. At the top and bottom of each page is a 'Get Features of Selected Sequences' button and when selected, a sorted, collated list of all the features within the sequences is displayed. These features can be selected using the check boxes and the type of FASTA output (DNA or protein) selected from the radio button at the bottom of the page. Finally the 'Get Fastas of Features' button at the bottom of the page will provide the sequences in a form that can be copied and pasted into further applications or saved to file. The process is illustrated for the coat proteins of members of the genus *Bromovirus* in Figure 1. The usefulness of the database lies in the ability to select multiple sequences of a species or genus and to obtain the sequences of one or more genes very rapidly. For example, in a recent test, a single complete sequence was selected from each species in the genus *Begomovirus*. From this point, the amino acid sequences of the coat proteins of each of the (139) species were obtained in ~70 s.

CLIENT SOFTWARE

As an additional method for accessing and analysing the sequence data, some client software has been written in Delphi that can be freely downloaded from the website to the user's PC. Three programs are currently available that use a similar interface to select the virus species, accession number(s) and sequence feature(s). Depending on the software, the program will then either (i) download the FASTA sequences (DNA or protein); (ii) calculate codon usage statistics; or (iii) predict transmembrane domains in the protein sequence using TMpred (3). Requests for data are made using http to the server and data is returned in XML format. While all of the data are centralized on the web database to make it easier to update, for speed of analysis and to reduce the server load, the subsequent calculations are handled by the client software on the PC and not by the PHP scripts. The client application makes http requests to the web server for information selected by the user and parses the XML data returned and formats it for display or processes it as required.

EXAMPLES OF USE

Sequence data from the database have proved valuable for a number of projects: (i) survey of codon usage bias amongst all plant viruses (4), (ii) two-way comparisons between comprehensive sets of sequences from the families *Flexiviridae* and *Potyviridae* that have helped inform taxonomy and clarify genus and species discrimination criteria (5,6), (iii) a survey and verification of the polyprotein cleavage sites within the family *Potyviridae* (7) together with dedicated web pages at http://www.rothamsted.bbsrc.ac.uk/ppi/links/pplinks/potycleavage/index.html.

CONCLUSIONS AND FUTURE PLANS

DPVweb and its associated tools provide an accurate, comprehensive and powerful resource for analysis of plant virus sequences. We plan to improve annotation of genes and proteins to provide more functional information where this is

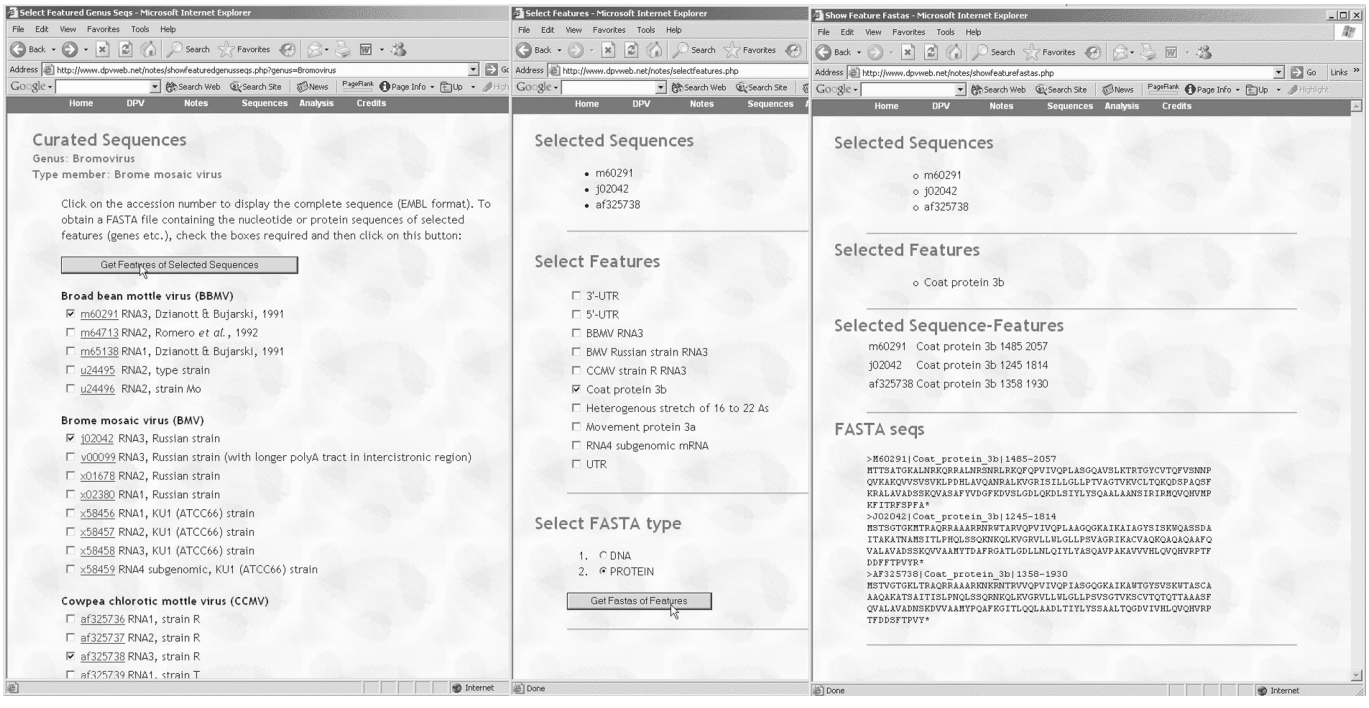

Figure 1. Sequence of screen shots from DPVweb.net showing (left) selection of curated sequences from species in the genus *Bromovirus*, (centre) selection of Coat protein feature and protein output and (right) the FASTA output of the coat protein genes.

available. We are also seeking collaborative opportunities to extend the usefulness of this project.

ACKNOWLEDGEMENTS

We thank our colleagues in the DPV project (Drs Hugh Barker, Tony Murant, Teifion Jones and David Robinson at Scottish Crop Research Institute and Phil Jones at Rothamsted Research) for their encouragement and suggestions. We thank the AAB for their support, in particular by funding the development of the standalone DPV program and for sponsoring the DPVweb internet site. Rothamsted Research receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom. Funding to pay the Open Access publication charges for this article was provided by Rothamsted Research Limited (http://www.rothamsted.ac.uk/).

Conflict of interest statement. None declared.

REFERENCES

1. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
2. Adams,M.J., Antoniw,J.F., Barker,H., Jones,A.T., Murant,A.F. and Robinson,D. (1998) *Descriptions of Plant Viruses on CD-ROM*. Association of Applied Biologists, Wellesbourne, Warwick, UK.
3. Hofmann,K. and Stoffel,W. (1993) TMbase—a database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler*, **374**, 166.
4. Adams,M.J. and Antoniw,J.F. (2004) Codon usage bias amongst plant viruses. *Arch. Virol*., **149**, 113–135.
5. Adams,M.J., Antoniw,J.F., Bar-Joseph,M., Brunt,A.A., Candresse,T., Foster,G.D., Martelli,G.P., Milne,R.G., Zavriev,S.K. and Fauquet,C.M. (2004) The new plant virus family *Flexiviridae* and assessment of molecular criteria for species demarcation. *Arch. Virol*., **149**, 1045–1060.
6. Adams,M.J., Antoniw,J.F. and Fauquet,C.M. (2005) Molecular criteria for genus and species discrimination within the family *Potyviridae*. *Arch. Virol*., **150**, 459–479.
7. Adams,M.J., Antoniw,J.F. and Beaudoin,F. (2005) Overview and analysis of the polyprotein cleavage sites in the family *Potyviridae*. *Mol. Plant Pathol*., **6**, 471–487.