Published online 7 December 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D339–D342***
doi:10.1093/nar/gkl1004

Genome Information Broker for Viruses (GIB-V): database for comparative analysis of virus genomes

Masaki Hirahata, Takashi Abe, Naoto Tanaka[1], Yoshikazu Kuwana[2], Yasumasa Shigemoto[2], Satoru Miyazaki[1], Yoshiyuki Suzuki and Hideaki Sugawara*

Center for Information Biology and DNA Data Bank of Japan National Institute of Genetics and Sokendai, Shizuoka, Japan, [1]Faculty of Pharmaceutical Sciences, Tokyo University of Science, Chiba, Japan and [2]Life Science Systems Division, Fujitsu Limited, Tokyo, Japan

Received September 15, 2006; Revised and Accepted October 27, 2006

ABSTRACT

Genome Information Broker for Viruses (GIB-V) is a comprehensive virus genome/segment database. We extracted 18 418 complete virus genomes/segments from the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/) by DNA Data Bank of Japan (DDBJ), EMBL and GenBank and stored them in our system. The list of registered viruses is arranged hierarchically according to taxonomy. Keyword searches can be performed for genome/segment data or biological features of any virus stored in GIB-V. GIB-V is equipped with a BLAST search function, and search results are displayed graphically or in list form. Moreover, the BLAST results can be used online with the ClustalW feature of the DDBJ. All available virus genome/segment data can be collected by the GIB-V download function. GIB-V can be accessed at no charge at http://gib-v.genes.nig.ac.jp/.

INTRODUCTION

Virus genome analysis has a long history. In 1977, Frederick Sanger successfully sequenced the entire genome of Phage ϕphiv;X174 (1). Since that time, a huge number of virus genomes have been sequenced. Virus genome sequences provide researchers with important information needed to analyze virus evolution, pathogenicity and diversity (2–4).

The International Nucleotide Sequence Database Collaboration (INSDC) makes a substantial amount of genomic data available in a public database. However, these genomic data are stored in the same database as general gene registrations and are not distinguished. In some cases, the specific prefix of the accession number (AE, AL, AP, BS, BX, CP, CR, CT, CU and CY) refers to a registration from a genome project (http://www.ddbj.nig.ac.jp/sub/prefix.html). However, for many genome data, the prefix is the same as that of general gene registrations, making it difficult to extract genomic data.

There are several virus genome databases such as DPVweb (5), HCVDB (6), VIDA (7) and VirGen (8). However, most of these are limited to specific groups of viruses. Thus, to fill the need for a comprehensive virus genome database, we constructed Genome Information Broker for Viruses (GIB-V), which includes all groups of viruses and is updated regularly with the release of our DNA Data Bank of Japan (DDBJ) data. GIB-V was created with the use of the Genome Information Browser (GIB) (9), an online microbial genome analysis system that we have developed.

GIB-V DATA SOURCE

The primary source of data for GIB-V was entries in the INSDC database. All of the virus genome data included in the virus division and phage division were targeted. In the INSDC database, each virus genome is registered in a single flat file if the virus does not have segments. If the virus genome has two or more segments, each segment is registered in a single flat file. Complete genome/segment data were identified by the description in the definition line of the flat file. We extracted genome records with the words ‘complete genome’ and without the words ‘Third Party Annotation (TPA)’ or ‘nearly complete’ in the definition line of the flat file. Segment records were extracted with a combination of the words ‘segment’ and ‘complete sequence’, and without the words ‘TPA’ or ‘nearly complete’ in the definition line of the flat file. These records were obtained from the virus division or the phage division of the DDBJ database. We extracted and stored 18 418 complete virus genome/segment data sets from release 66, which was the latest DDBJ data released at the time of the initial development of GIB-V (Table 1). DDBJ releases new data four times per year. GIB-V is updated at the time of each release.

SYSTEM ARCHITECTURE OF GIB-V

The system architecture and the data flow of GIB-V are depicted in Figure 1. The World Wide Web server for

*To whom correspondence should be addressed. Tel: +81 55 981 6895; Fax: +81 55 981 6896; Email: hsugawar@genes.nig.ac.jp

GIB-V is Apache (http://www.apache.org/) with Hypertext Preprocessor (PHP, http://www.php.net). PHP is a server-side HTML-embedded scripting language and is able to dynamically generate HTML page contents. GIB-V uses PostgreSQL (http://www.postgresql.org/) as its relational database management system and is distributed over multiple PC Linux platforms.

The genome/segment data are extracted from the DDBJ database and stored in the relational database server. The nucleotide sequence data and the protein sequence data are also extracted from the flat files and stored to the homology search server.

In the original GIB, Common Object Request Broker Architecture (CORBA) was used for internal communication. However, CORBA was not needed for GIB-V because of remarkable advances in hardware performance that solved the problems associated with managing the large quantity of data.

Table 1. Number of registered genomes/segments in GIB-V

Genome Name of virus	Number	Segment Name of virus	Number
Hepatitis B virus	843	Influenza A virus	11 068
Human immunodeficiency virus 1	784	East African cassava mosaic virus	96
JC polyomavirus	371	Bluetongue virus	51
Porcine circovirus 2	171	Cucumber mosaic virus	35
Dengue virus	160	Cypovirus 1	34
Foot-and-mouth disease virus	129	Tomato spotted wilt virus	34
SARS coronavirus	125	Cotesia congregata bracovirus	30
Hepatitis C virus	113	Rotavirus A	30
Hop stunt viroid	88	Crimean-Congo hemorrhagic fever virus	28
Peach latent mosaic viroid	80	Rice stripe virus	28

This table indicates the top 10 registered viruses. Please refer to web page statistics for complete data.

FUNCTIONS OF GIB-V

The functions of GIB-V are described below. Most of these functions are accessed from the main menu bar displayed in the Web page.

Genome list

The complete list of virus data stored in GIB-V can be viewed by clicking ‘GENOME LIST’ on the main menu bar. Viruses are displayed at the family level on an initial screen. The list of genera is displayed by clicking on each family name; species are displayed by clicking on genus names, and organisms are displayed by clicking on species names. The accession number of entries in the INSDC database is displayed under the name of the organism, and the ‘Virus Information Page’ is opened by clicking on the accession number. The ‘Virus Information Page’ provides access to information about the individual virus genome/segment. When two or more genomes/segments are registered at the same virus name, each genome/segment can be accessed by selecting the related accession number in the pull-down menu. From the ‘Virus Information Page’, it is possible to move to three kinds of genome/segment information pages: (i) ‘Feature View,’ which displays a graphic of the specified region; (ii) ‘Feature List,’ which provides a tabular listing of features included in the genome/segment; and (iii) ‘Nuc Sequence,’ which indicates the nucleic acid sequence of the genome/segment. In the ‘Feature View,’ each open reading frame (ORF) is indicated by a white bar in the regional chart displayed with the G+C% graph (GC plot), and it is possible to jump to the ‘Feature Information’ page by clicking the specified ORF bar. From the ‘Feature List,’ the user has the option to display the whole region or a specified region of the genome/segment. The name of the species, the start and end points of its location and its product name are included in the ‘Feature List.’ It is possible to jump to

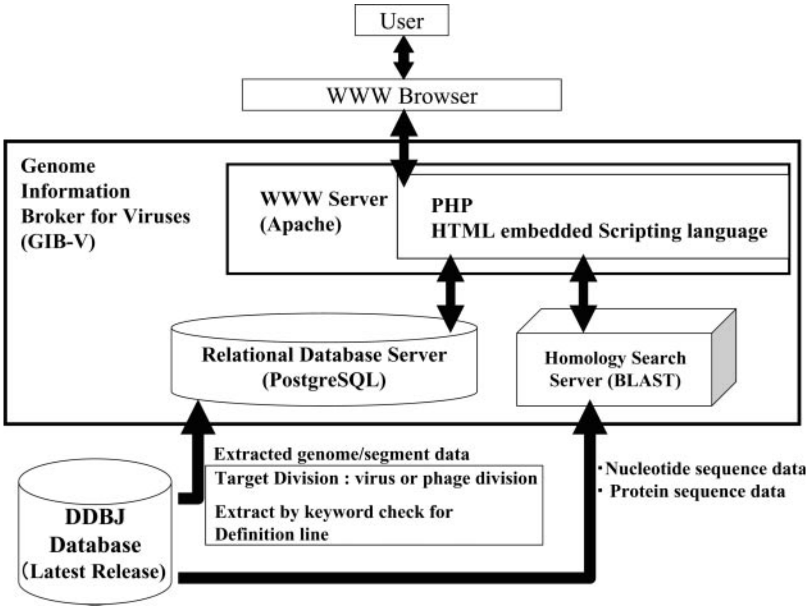
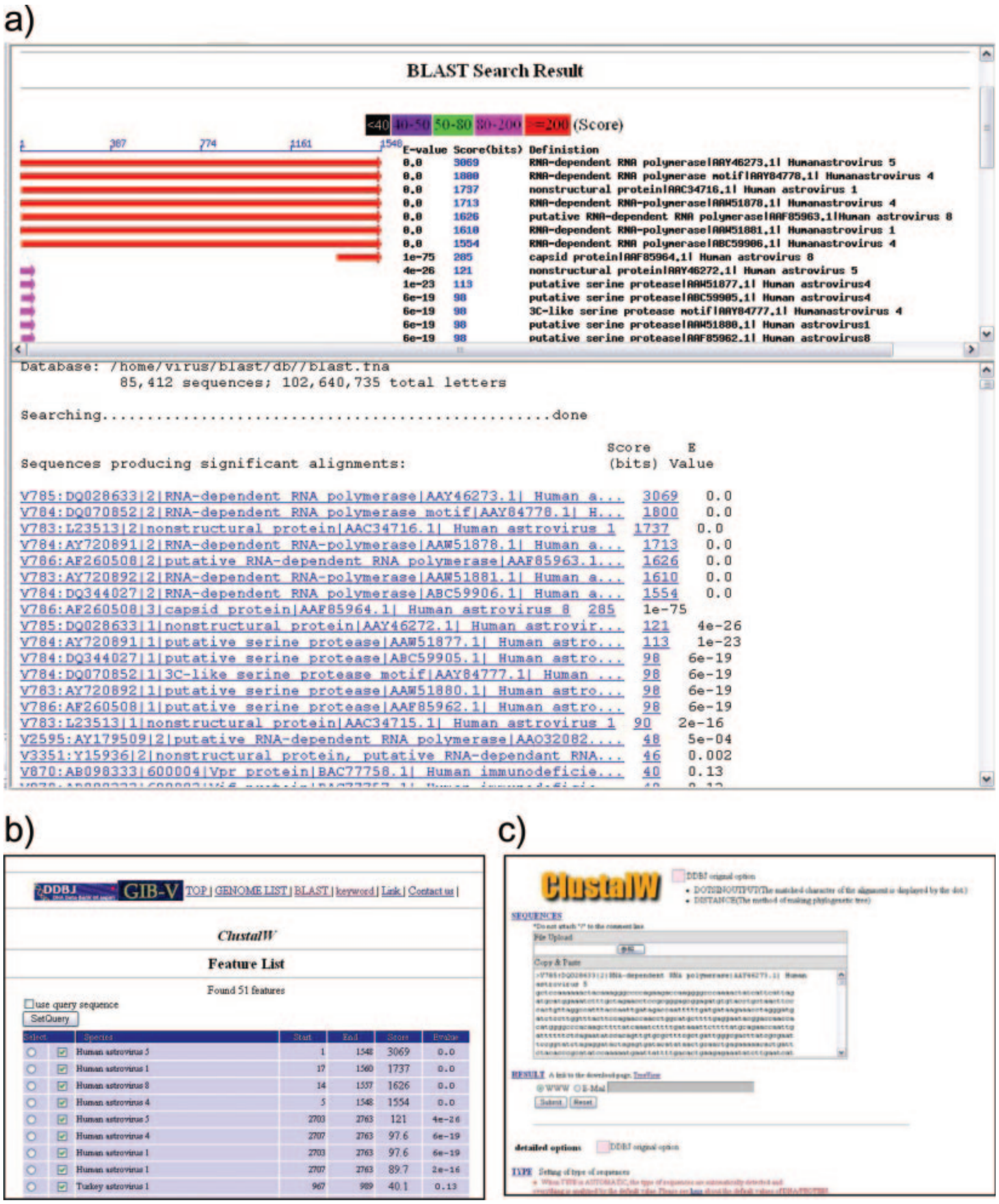

User
WWW Browser
Genome Information Broker for Viruses (GIB-V)
WWW Server (Apache)
PHP
HTML embedded Scripting language
Relational Database Server (PostgreSQL)
Homology Search Server (BLAST)
Extracted genome/segment data
Target Division : virus or phage division
Extract by keyword check for Definition line
DDBJ Database (Latest Release)
・Nucleotide sequence data
・Protein sequence data

Figure 1. Schema of Genome Information Broker for Viruses (GIB-V). This schema shows the system architecture and the data flow of GIB-V. The hardware components of GIB-V are a World Wide Web server, relational database server and homology search server.



```
E-value Score(bits) Definition
0.0     3069        RNA-dependent RNA polymerase|AAY46273.1| Humanastrovirus 5
0.0     1800        RNA-dependent RNA polymerase motif|AAY84778.1| Humanastrovirus 4
0.0     1737        nonstructural protein|AAC34716.1| Human astrovirus 1
0.0     1713        RNA-dependent RNA-polymerase|AAW51878.1| Humanastrovirus 4
0.0     1626        putative RNA-dependent RNA polymerase|AAF85963.1|Human astrovirus 8
0.0     1610        RNA-dependent RNA-polymerase|AAW51881.1| Humanastrovirus 1
0.0     1554        RNA-dependent RNA polymerase|ABC59906.1| Humanastrovirus 4
1e-75   285         capsid protein|AAF85964.1| Human astrovirus 8
4e-26   121         nonstructural protein|AAY46272.1| Human astrovirus 5
1e-23   113         putative serine protease|AAW51877.1| Human astrovirus4
6e-19   98          putative serine protease|ABC59905.1| Human astrovirus4
6e-19   98          3C-like serine protease motif|AAY84777.1| Humanastrovirus 4
6e-19   98          putative serine protease|AAW51880.1| Human astrovirus1
6e-19   98          putative serine protease|AAF85962.1| Human astrovirus8
```

```
Database: /home/virus/blast/db//blast.fna
           85,412 sequences; 102,640,735 total letters

Searching..................................................done

                                                               Score    E
Sequences producing significant alignments:                    (bits) Value

V785:DQ028633|2|RNA-dependent RNA polymerase|AAY46273.1| Human a...  3069    0.0
V784:DQ070852|2|RNA-dependent RNA polymerase motif|AAY84778.1| H...  1800    0.0
V783:L23513|2|nonstructural protein|AAC34716.1| Human astrovirus 1  1737    0.0
V784:AY720891|2|RNA-dependent RNA-polymerase|AAW51878.1| Human a...  1713    0.0
V786:AF260508|2|putative RNA-dependent RNA polymerase|AAF85963.1...  1626    0.0
V783:AY720892|2|RNA-dependent RNA-polymerase|AAW51881.1| Human a...  1610    0.0
V784:DQ344027|2|RNA-dependent RNA polymerase|ABC59906.1| Human a...  1554    0.0
V786:AF260508|3|capsid protein|AAF85964.1| Human astrovirus 8  285    1e-75
V785:DQ028633|1|nonstructural protein|AAY46272.1| Human astrovir...  121    4e-26
V784:AY720891|1|putative serine protease|AAW51877.1| Human astro...  113    1e-23
V784:DQ344027|1|putative serine protease|ABC59905.1| Human astro...  98    6e-19
V784:DQ070852|1|3C-like serine protease motif|AAY84777.1| Human ...  98    6e-19
V783:AY720892|1|putative serine protease|AAW51880.1| Human astro...  98    6e-19
V786:AF260508|1|putative serine protease|AAF85962.1| Human astro...  98    6e-19
V783:L23513|1|nonstructural protein|AAC34715.1| Human astrovirus 1  90    2e-16
V2595:AY179509|2|putative RNA-dependent RNA polymerase|AAO32082....  48    5e-04
V3351:Y15936|2|nonstructural protein, putative RNA-dependant RNA...  46    0.002
V870:AB098333|600004|Vpr protein|BAC77758.1| Human immunodeficie...  40    0.13
```

ClustalW

Feature List

Found 51 features

☐ use query sequence

Select		Species	Start	End	Score	Evalue
○	☑	Human astrovirus 5	1	1548	3069	0.0
○	☑	Human astrovirus 1	17	1560	1737	0.0
○	☑	Human astrovirus 8	14	1557	1626	0.0
○	☑	Human astrovirus 4	5	1548	1554	0.0
○	☑	Human astrovirus 5	2703	2763	121	4e-26
○	☑	Human astrovirus 4	2707	2763	97.6	6e-19
○	☑	Human astrovirus 1	2703	2763	97.6	6e-19
○	☑	Human astrovirus 1	2707	2763	89.7	2e-16
○	☑	Turkey astrovirus 1	967	989	40.1	0.13

Figure 2. Example of BLAST search results. blastp was executed for the 'All family' sequence. The query is an RNA-dependent RNA-polymerase of Human astrovirus. (**a**) Search result as it appears with 'Graphic View' indicated. The upper part of the page shows the result displayed graphically, and the lower part shows the result in text form. (**b**) Search result as it appears with 'Feature List' indicated. Clicking the 'Set Query' button activates ClustalW and sends sequence data automatically. (**c**) The ClustalW page. Data has been submitted from the Feature List.

'Feature View' or 'Features Information' page from the list. The nucleic acid sequence or amino acid sequence of each ORF in the list can be downloaded. The 'Nuc Sequence' page displays the nucleic acid sequence of the whole region or a specified region in a text format. The 'Feature Information' page displays the feature/qualifier information for each ORF. Both the amino acid and nucleic acid sequence are displayed with the feature/qualifier information. The upper or lower stream of the nucleic acid sequence can be displayed.

For the viroid genomes, a graphical display is not provided because it does not work effectively without the ORF. However, the user can download sequence data from the download page.

BLAST search

GIB-V can execute a BLAST search (10) for ORF sequences of genomes/segments. The user can select from blastn, blastp, blastx and tblastx search options. This search is executed in two steps. The BLAST search page is opened by clicking on 'BLAST' in the main menu bar, and a virus is selected from the list on the 'BLAST Search-Subject selection page.' A single family or all families can be selected. The

Subject can be narrowed down to the genus level of the family when a single family is selected. A BLAST program and other options are then selected, and the query is submitted.

The user can select to have BLAST Search results presented as a graphical display (Graphic View) or a list display (Feature List). In the 'Graphic View,' alignment information is displayed both graphically and as text. The position of the hit area in relation to the query is displayed graphically, and the score is displayed by color (Figure 2a). In the 'Feature List,' a summary of the results is displayed as a list (Figure 2b). After a target candidate ORF is selected, it is possible to jump directly from the list to the ClustalW feature of the DDBJ (Figure 2c).

Keyword search

To perform a keyword search, the user clicks 'KEYWORD' on the main menu bar. The user can find either viruses or ORFs by the combinations of multiple keywords up to 5 terms. INSDC accession number, virus name, country, note, isolate, segment, serotype, specific host and strain are usable for viruses search. These categories are the types of biological information included in genome/segment flat files from the INSDC. The accession number of the target genome/segment can be reached easily from the list of search results. Selected genomes/segments in the list can be downloaded with a FASTA or flat file format in a lump. As to ORFs search, db_xref (pointer to related information in another database), EC number, function, gene name, product name, protein ID, bound moiety (molecule/complex that may bind to the given feature), phenotype, location and other qualifiers defined by INSDC can be specified. Target ORF information can be accessed from the search results. The retrieved ORF can be displayed by nucleic acid or amino acid sequence in a text format.

Download function

Genome/segment data are downloadable from GIB-V. On the 'Genome List' page, there are checkboxes in front of each taxonomic name lower than the family level. The user clicks the target checkbox and download button to download the specified genome/segment data in flat file form. The download page also can be entered from each 'Virus Information' page. From the download page, four types of data can be downloaded: a whole genome flat file (DDBJ format); whole genome sequence (FASTA format); nucleotide sequences of all features (FASTA format); and amino acid sequences of the cording sequence (CDS) (FASTA format). When two or more genomes/segments are registered for the same virus name, they can be downloaded from the same page.

FUTURE PLANS

We have developed G-InforBIO system, a tool for genome data management and sequence analysis (http://wdcm.nig.ac.jp/inforbio/G-InforBIO/download.html) (11). It is equipped with a variety of software functions and can perform seamless analysis of multiple genomes. It was modified to adapt the GIB-V output data form. We are going to reinforce the cooperation of GIB-V and G-InforBIO to improve the smoothness of data communication between them, and expand the genome comparison environment. We will consider the enrichment of the contents with the improvement of data extraction process and the addition of incidental information.

ACCESS TO THE DATABASE

GIB-V can be accessed from http://gib-v.genes.nig.ac.jp/. Registration is not necessary, and use of the database is free.

ACKNOWLEDGEMENTS

The development of GIB-V was supported partly by the Institute for Bioinformatics Research and Development of Japan Science and Technology Agency. Funding to pay the Open Access publication charges for this article was provided by the Institute for Bioinformatics Research and Development of Japan Science and Technology Agency.

Conflict of interest statement. None declared.

REFERENCES

1. Sanger,F., Air,G.M., Barrell,B.G., Brown,N.L., Coulson,A.R., Fiddes,C.A., Hutchison,C.A., Slocombe,P.M. and Smith,M. (1977) Nucleotide sequence of bacteriophage phiX174 DNA. *Nature*, **265**, 687–695.
2. Vijgen,L., Keyaerts,E., Lemey,P., Maes,P., Van Reeth,K., Nauwynck,H., Pensaert,M. and Van Ranst,M. (2006) Evolutionary history of the closely related group 2 coronaviruses: porcine hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43. *J. Virol*., **80**, 7270–7274.
3. Zhou,J.Y., Shen,H.G., Chen,H.X., Tong,G.Z., Liao,M., Yang,H.C. and Liu,J.X. (2006) Characterization of a highly pathogenic H5N1 influenza virus derived from bar-headed geese in China. *J. Gen. Virol*., **87**, 1823–1833.
4. Culley,A.I., Lang,A.S. and Suttle,C.A. (2006) Metagenomic analysis of coastal RNA virus communities. *Science*, **312**, 1795–1798.
5. Adams,M.J. and Antoniw,J.F. (2006) DPVweb: a comprehensive database of plant and fungal virus genes and genomes. *Nucleic Acids Res*., **34**, D382–D385.
6. Combet,C., Penin,F., Geourjon,C. and Deleage,G. (2004) HCVDB: hepatitis C virus sequences database. *Appl. Bioinformatics*, **3**, 237–240.
7. 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.
8. Kulkarni-Kale,U., Bhosle,S., Manjari,G.S. and Kolaskar,A.S. (2004) VirGen: a comprehensive viral genome resource. *Nucleic Acids Res*., **32**, D289–D292.
9. Fumoto,M., Miyazaki,S. and Sugawara,H. (2002) Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. *Nucleic Acids Res*., **30**, 66–68.
10. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
11. Tanaka,N., Abe,T., Miyazaki,S. and Sugawara,H. (2006) G-InforBIO: integrated system for microbial genomics. *BMC bioinformatics*, **7**, 368.