Poxvirus Bioinformatics Resource Center: a comprehensive Poxviridae informational and analytical resource

Elliot J. Lefkowitz*, Chris Upton¹, Shankar S. Changayil, Charles Buck², Paula Traktman³ and R. Mark L. Buller⁴

Department of Microbiology, University of Alabama at Birmingham, BBRB 276/11; 1530 3rd Avenue S., Birmingham, AL 35294-2170, USA, ¹Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada V8W 2Y2, ²Virology Collection, ATCC, Manassas, VA 20108, USA, ³Department of Microbiology and Molecular Genetics, Medical College of Wisconsin, Room 273–BSB, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA and ⁴Department of Molecular Microbiology and Immunology, St Louis University Health Sciences Center, 1402 South Grand Boulevard, St Louis, MO 63104, USA

Received August 12, 2004; Revised and Accepted October 18, 2004

ABSTRACT

The Poxvirus Bioinformatics Resource Center (PBRC) has been established to provide informational and analytical resources to the scientific community to aid research directed at providing a better understanding of the Poxviridae family of viruses. The PBRC was specifically established as the result of the concern that variola virus, the causative agent of smallpox, as well as related viruses, might be utilized as biological weapons. In addition, the PBRC supports research on poxviruses that might be considered new and emerging infectious agents such as monkeypox virus. The PBRC consists of a relational database and web application that supports the data storage, annotation, analysis and information exchange goals of the project. The current release consists of over 35 complete genomic sequences of various genera, species and strains of viruses from the Poxviridae family. Sequence and annotation information for these viruses has been obtained from sequences publicly available from GenBank as well as sequences not yet deposited in GenBank that have been obtained from ongoing sequencing projects. In addition to sequence data, the PBRC provides comprehensive annotation and curation of virus genes; analytical tools to aid in the understanding of the available sequence data, including tools for the comparative analysis of different virus isolates; and visualization tools to help better display the results of various analyses. The PBRC represents the initial development of what will become a more comprehensive Viral Bioinformatics Resource Center for Biodefense that will be one of the National Institute of Allergy and Infectious Diseases’ ‘Bioinformatics Resource Centers for Biodefense and Emerging or Re-Emerging Infectious Diseases’. The PBRC website is available at http://www.poxvirus.org.

INTRODUCTION

An effective response to the use of biological organisms as agents of terrorism or warfare, or to the emergence of new infectious diseases requires a multi-disciplinary effort involving various agencies at the local, state and federal levels including public health officials, hospital personnel, epidemiologists and the military. In addition to the public health response, a concerted research effort is necessary to better detect, understand and respond to these threats. Such research requires development of environmental detectors and clinical diagnostic aids to provide us with rapid warning in the event of an outbreak as well as development of vaccines to prevent infection and antiviral or antibacterial drugs to cure infection. These efforts require a comprehensive biological understanding of potential threat agents, including their molecular biology, genetics, pathogenicity, epidemiology and evolution. The National Institute of Allergy and Infectious Diseases as well as the US Centers for Disease Control and Prevention maintain a list of priority pathogens that are considered potential biothreat agents and/or are microbes that appear to be new or
reemerging pathogens (http://www.bt.cdc.gov/agent/agentlist-category.asp and http://www.niaid.nih.gov/biodefense/bandc_priority.htm). Variola virus, the causative agent of smallpox and a member of the Poxviridae family of viruses, has perhaps the greatest potential for use as a bio-weapon and is one of the Category A pathogens on these priority pathogen lists (1). In addition, monkeypox virus, a member of the orthopoxvirus genus that includes variola virus, has caused a number of disease outbreaks in recent years, including outbreaks in North America resulting from the importation of rodents from Africa intended to be sold as pets (2,3).

The use of high-throughput DNA sequencing techniques as well as other large-scale ‘Systems Biology’ technologies have led to an unprecedented increase in the amount of available data. Therefore, one overarching necessity in research efforts directed at providing a better understanding of priority pathogens is the need to collect, manage, describe, analyze and publicize the vast amounts of information generated by modern, high-throughput biological research. Therefore, the goal of the Poxvirus Bioinformatics Resource Center (PBRC) is to organize all available information on virus genetics thus aiding research efforts towards increasing our knowledge of virus replication and virus–host interaction on a gene-by-gene and whole genome basis. In addition, the PBRC is expanding on available knowledge by developing and utilizing analysis tools that can further probe the information contained in the genome and gene sequences of these organisms. Since our goal is to establish an information resource to support research efforts by the scientific community, we are also soliciting input from that community to ensure the completeness and, above all, the accuracy of the information being provided and to ensure that the software tools provided and in development reflect the needs of the different research groups using these resources.

WHAT IS THE POXVIRUS BIOINFORMATICS RESOURCE CENTER?

The PBRC represents a cooperative endeavor with collaborations between the University of Alabama at Birmingham, the University of Victoria, the Medical College of Wisconsin, St Louis University and the American Type Culture Collection (ATCC). The PBRC consortium has established a database of all available completely sequenced poxvirus genomes. This database includes information on every predicted gene that may be coded for by these genomes, as well as descriptive annotations of the physical and functional properties of each gene based on computer predictions. Currently, only complete genomes are included in the database. In future releases we plan to include as available, incomplete genomes as well as coding sequences from virus strains not represented by complete genomic sequences. We are also compiling a comprehensive gene-by-gene curation that is linked to each individual gene record. In addition, the PBRC provides a variety of analytical information and analysis tools that can be used to mine the genomic data. These tools include sequence homology searches, a database of functional domains, a database of poxvirus gene orthologs and web-based visualization tools to allow for customized displays of much of this information. A major goal has been to develop new software packages for the analysis of viral genomes. This aspect of the project involves designing new software tools that permit researchers to interact with and manipulate complete poxvirus genomes and families of poxvirus protein orthologs. This vastly speeds up the analysis process and provides information about these viruses from comparative analyses that otherwise would be almost impossible to obtain. Some of these tools have been described previously in the literature and therefore more substantial descriptions are already available [Poxvirus Orthologous Clusters (4,5), Viral Genome Organizer (6), Viral Genome Database (7), JDotter (8) and Base-By-Base (9)].

DATABASE DESCRIPTION

The basic PBRC web portal accesses genomic, annotative and analytical information from a Microsoft SQL Server database. A companion MySQL database provides data to a number of java-based analytical tools. The database schema used for the PBRC originally was developed to accommodate bacterial genomes and their genes (protein and RNA) (10,11) and needed only slight modification to support the storage of information on poxvirus genomes and genes (12). The PBRC data schema provides tables to store sequences and basic sequence annotations, human-annotated curation records and tables for the storage of basic analytical information such as the results of BLAST searches, functional motifs and biophysical properties. The database supports all of the web-based query tools and also serves as the data source for all of our web-based analytical tools.

GENOME DATA ACQUISITION

Genomic data are obtained from GenBank (13), other publicly available databases and websites, as well as from ongoing sequencing projects. The current PBRC release consists of over 35 complete genomes from the Poxviridae family of viruses. If available from the GenBank record or from the sequencing laboratory, the predicted gene set of any one genome is stored in the PBRC database and used as the initial starting point for annotation and curation of each genome. If a predicted gene set is not available, we perform our own gene prediction using a combination of gene predictive tools that start with open reading frames, and then include among others, promoter prediction, presence of functional motifs and assignment to orthologous protein clusters. Our experience is that the methods used for the prediction of protein coding genes for each poxvirus genome present in GenBank has in many cases been derived using different parameters. Therefore, we are in the process of developing a more consistent method for the prediction of poxvirus genes and will utilize this gene prediction pipeline to re-assess the gene set for each available genome.

Each virus genome is categorized according to its genus, species and strain designation as determined by the International Committee for the Taxonomy of Viruses (ICTV) (http://www.ncbi.nlm.nih.gov/ICTVdb/) (14,15). In addition to providing the nomenclature available from the GenBank record for each gene in any one genome, we also provide
our own gene designation which is a combination of the ICTV-approved species name, the strain or isolate name and a numerical designation for each gene that starts at number 1 for the left-most gene, and is incremented by one for each subsequent gene as determined by its genome position.

**GENOME ANNOTATION AND CURATION**

For each gene in the PBRC database, we provide an automated, computer-driven annotation that gathers as much basic descriptive information as possible about a gene, basic analysis of its nucleotide or amino acid sequence, and the results of sequence similarity searches to look for common patterns or features that might be characteristic of its function. The annotation process starts with the GenBank record and includes the descriptive information, literature references and any other information provided in that record. This information populates the initial descriptive fields of our database. Following this automated annotation process, a manual, human-directed curation of each gene record is undertaken. During this curation process, a researcher reviews the annotation record, all available literature references and any unpublished information as available. This collection of empirically derived properties for the protein in question provides what might be considered a mini-review of the biology of the gene being studied. The broad types of information that are provided during the curation process include protein properties such as molecular weight and pI; post-translational processing; the availability of custom reagents such as clones, antibodies and mutants; functional descriptions [including Gene Ontology designations (16)]; and literature summaries. Evidence codes are provided that explicitly state the nature and source of each piece of information along with the appropriate literature references. A series of web forms assist in this process that provides a distinct set of informational fields to be filled in, and enforces use of a controlled vocabulary to fully describe each gene. The results of the curation process are stored in our SQL Server database and form a Poxvirus Knowledge Database that is available and searchable from the PBRC website.

**WEBSITE**

The PBRC website is provided by a Microsoft Windows 2003 server running Microsoft Internet Information Services (IIS). The user interface is provided through a combination of Active Server Pages running server-side Visual Basic script, client-side JavaScript and HTML. SQL Server database access is provided through Microsoft ActiveX Data Objects.

The starting point for access to all databases and tools available from the PBRC is the PBRC home page at http://www.poxvirus.org/. A menu is displayed along the top of the page and appropriate clicking on a menu link brings up a submenu that provides access to individual web query tools and/or applications. The Data menu provides access to search forms that provide user-specified queries for genome, gene and sequence data (Figures 1 and 2). Analytical tools are provided through another series of submenus and provide access to a series of analytical and visualization tools (Figure 3). In addition to the main menu and context-sensitive

![Figure 1. PBRC genomic sequence web pages. Screen shots of the PBRC genome list, and the genome map for *Variola major* virus strain Bangladesh.](image-url)
submenu, a menu near the bottom of the page provides access to PBRC organizational information such as descriptions of the people and teams involved in the work, acknowledgements and a form to provide user feedback. At the very bottom of the form we provide a Google-supported search form.

In general, available web pages can be categorized as follows:

(i) *Informational*: Static text such as help files, biographical information or meeting announcements.

(ii) *Forms*: Pages that provide user input for data queries.

**Figure 2.** PBRC gene search and gene record web pages. Screen shots of the PBRC gene search form, a listing of gene search results and an individual gene record.
(iii) **Tabular results**: Pages displaying data in tables.

(iv) **Graphical results**: Pages displaying data results as figures.

(v) **Links**: Pages providing access to other pages.

(vi) **Applets**: Small Java-based applications for data analysis and visualization that are available from within a user’s web browser.

(vii) **Applications**: More comprehensive Java-based analytical applications that can be run from a web page using Java Web Start, or downloaded to a user’s workstation and run as a stand-alone application.

**ANALYTICAL TOOLS**

BLAST similarity searches (17) are available that provide standard web-based BLAST output as well as a web-based version of our BLAST parsing and visualization application, XS-BLAST (XML-SQL BLAST). XS-BLAST utilizes the XML-output option of the NCBI BLAST executable, and parses the results storing them in our SQL Server database. The result set is then provided to the user as an HTML table, as well as from a Java-based graphical visualization tool (Figure 3). Both BLAST searches are run locally on a Sun Solaris server and provide several customized databases for searching. These include all complete genomic Poxviridae nucleic acid sequences, all predicted Poxviridae protein sequences and all available Poxviridae nucleotide sequences extracted from GenBank.

Comparative analysis of the coding potential of all Poxviridae genomes is available as both pairwise and multi-genome comparisons. Both tabular listings of orthologous protein sets are provided as well as a graphical gene synten plot of shared orthologs between any two genomes (Figure 3). Ortholog determination is based on an all versus all BLAST search of all poxvirus proteins in the PBRC database.

Figure 3. PBRC BLAST search and gene syteny analytical tool web pages. Screen shots of XS-BLAST tabular and graphical BLAST search results and a gene synteny plot.
REFERENCES

1. Henderson, D.A., Inglesby, T.V., Bartlett, J.G., Ascher, M.S., Eitzen, E., Jhariing, P.B., Bauer, J., Layton, M., McDade, J., Osterholm, M.T. et al. (1999) Smallpox as a biological weapon: medical and public health management. Working Group on Civilian Biodefense. J. Am. Med. Assoc., 281, 2127–2137.

2. Guarnier, J., Johnson, B.J., Paddock, C.D., Shieh, W.J., Goldsmith, C.S., Reynolds, M.G., Damron, I.K., Regner, R.L. and Zaki, S.R. (2004) Monkeypox transmission and pathogenesis in prairie dogs. Emerg. Infect. Dis., 10, 426–431.

3. Enserink, M. (2003) Infectious diseases. U.S. monkeypox outbreak traced to Wisconsin pet dealer. Science, 300, 1639.

4. Upton, C., Slack, S., Hunter, A.W., Ehlers, A. and Roper, R.L. (2003) Poxvirus Orthologous Clusters (POCs): toward defining the minimum essential poxvirus genome. J. Virol., 77, 7590–7600.

5. Ehlers, A., Osborne, J., Slack, S., Roper, R.L. and Upton, C. (2002) Poxvirus Orthologous Clusters (POCs). Bioinformatics, 18, 1544–1545.

6. Upton, C., Hogg, D., Perrin, D., Boone, M. and Harris, N.L. (2000) Viral genome organizer: a system for analyzing complete viral genomes. Virus Res., 70, 55–64.

7. Hiscock, D. and Upton, C. (2000) Viral Genome Database: storing and analyzing genes and proteins from complete viral genomes. Bioinformatics, 16, 484–485.

8. Brodie, R., Roper, R.L. and Upton, C. (2004) JDotter: a Java interface to multiple dotplots generated by Dotter. Bioinformatics, 20, 279–281.

9. Brodie, R., Smith, A.J., Roper, R.L., Tcherepanov, V. and Upton, C. (2004) Base-By-Base: single nucleotide-level analysis of whole viral genome alignments. BMC Bioinformatics, 5, 96.

10. Hoskins, J., Albom, W.E., Jr, Arnold, J., Blassczak, L.C., Burgett, S., DeHoff, B.S., Estrem, S.T., Fritz, L., Fu, D.J., Fuller, W. et al. (2001) Genome of the bacterium Streptococcus pneumoniae strain R6. J. Bacteriol., 183, 5709–5717.

11. Glass, J.J., Lefkowitz, E.J., Glass, J.S., Heiner, C.R., Chen, E.Y. and Cassell, G.H. (2000) The complete sequence of the mucosal pathogen Ureaplasma urealyticum. Nature, 407, 757–762.

12. Chen, N., Danila, M.I., Feng, Z., Buller, R.M.L., Wang, C., Han, X., Lefkowitz, E. and Upton, C. (2003) The genomic sequence of Entomel lia virus, the causative agent of mousepox. Virology, 317, 165–186.

13. 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. et al. (2003) Database resources of the National Center for Biotechnology. Nucleic Acids Res., 31, 28–33.

14. Ball, L.A. and Mayo, M.A. (2004) Virology Division News: report from the S3rd Meeting of the ICTV Executive Committee. Arch. Virol., 149, 1259–1263.

15. Fauquet, C.M. and Mayo, M.A. (2001) The 7th ICTV report. Arch. Virol., 146, 189–194.

16. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res., 32, D258–D261.

17. Altschul, S.F., Madden, T.L., Schaffer, 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.