BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 12 2009, pages 1568–1569
doi:10.1093/bioinformatics/btp258

Databases and ontologies

ProtVirDB: a database of protozoan virulent proteins

Jayashree Ramana and Dinesh Gupta*

Structural and Computational Biology Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhi 110067, India

Received on February 13, 2009; revised on March 23, 2009; accepted on April 8, 2009

Advance Access publication April 15, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Summary: ProtVirDB is a comprehensive and user-friendly web-based knowledgebase of virulent proteins belonging to protozoan species. The database will facilitate research and provide an integrated platform for comparative studies of virulent proteins in different parasitic protozoans and organize them under a unifying classification schema with functional categories. Remarkably, one-third of the protein sequences in the database showed presence of either mono- or hetero-repeats, or both concomitantly—hence reiterating the importance of repeats in parasite virulence mechanisms. A number of useful bioinformatics tools including BLAST and tools for phylogenetic analysis are integrated with the database. With the rapidly burgeoning interest in the pathogenesis mechanisms of protozoans and ongoing genome sequencing projects, we anticipate that the database will be a useful tool for the research community.

Availability: http://bioinfo.icgeb.res.in/protvirdb

Contact: dinesh@icgeb.res.in

Supplementary information: Supplementary data are available at *Bioinformatics* online

1 INTRODUCTION

Virulent proteins are an important class of proteins enabling pathogens to evade host immune mechanisms to cause disease in the host. There is an ever-growing interest to identify novel virulent proteins in variety of pathogens in order to counter growing drug resistance and to develop novel vaccines. Well-defined classes for bacterial virulent proteins (Prescott *et al.,* 1999) have been described; however, no such classification is available for protozoan virulent proteins. There are databases of bacterial virulent proteins like VFDB, PRINTS and MVirDB (Zhou *et al*., 2006); however, there is no report of any such database for protozoans.

In this work, our main objective was to develop a unified information portal for the researchers interested in a panoramic or in-depth view of the virulent proteins in a parasite or in comparison with other parasites. We have attempted a function-based classification for these proteins. However, the delineation amongst different categories is rather vague, for example adhesion and invasion (e.g. CSP protein from *Plasmodium falciparum*), and is for the purpose of a broad outline.

*To whom correspondence should be addressed.

2 DATABASE CONTENT

ProtVirDB is a non-redundant database currently holding a cumulative collection of 345 unique virulent proteins (however, number of total entries is 1775, as the database contains several polymorphic proteins) from 12 important parasitic protozoans (Supplementary Material 1). The database entries were manually curated from bibliographic (PubMed) and sequence (GenBank, RefSeq and SwissProt) databases (Fig. 1). Based on the currently available literature, each protein was allotted to one of the following functional categories: Adhesin, Invasion, Establishment (within the host, i.e. proteins involved in nutrient acquisition or evasion of host immunity), Proteases, Cysteine proteases, Heat shock proteins and Others. Cysteine proteases serve a multitude of roles like cytoadherence, invasion, etc., so an exclusive category has been devoted to these.

3 DATABASE ARCHITECTURE AND DATA RETRIEVAL

ProtVirDB is implemented as a MySQL database (Supplementary Material 2), PHP is used to connect the database and dynamically generate user-friendly HTML front-end queries, using Apache web server. The web interface query form allows users to selectively retrieve a table enlisting details (functional category and a brief

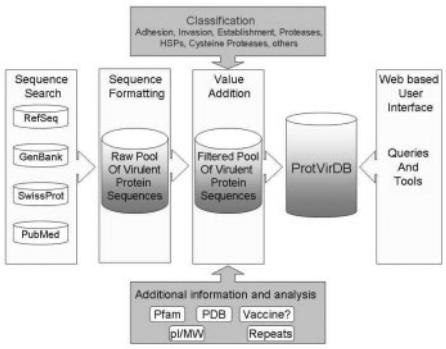

Fig. 1. ProtVirDB database schema. The protein sequences were collected by keyword search from different databases and then filtered by retrieving related articles from PubMed. Value addition included classification of filtered pool sequences into different categories and additional analysis (like Pfam domains, pI/Mw, PDB code, vaccine potential and presence of amino acid repeats). The web-based user interface integrated with powerful bioinformatics tools facilitates database query and analysis.

description) of virulent proteins from one or more organisms within a single or multiple functional categories. Users can selectively download the proteins of interest as an excel table or FASTA file. Each protein is linked to additional information comprising its molecular weight, pI, PDB code, Pfam domains (Finn *et al*., 2008), amino acid repeats and PubMed links. Additional PubMed links for verified or predicted vaccine or immunotherapeutic targets are included. Links have been provided for TDR drug target database entries (Aguero *et al.*, 2008). Alternatively, the database can be queried with user-defined keywords, with accession number or molecular weight combined with the filter based on organism name. Links to several web-based utilities for calculation of antigenic index and epitope prediction are provided.

4 INTEGRATED WEB-BASED TOOLS

ProtVirDB is integrated with several useful tools to facilitate sequence retrieval and analysis. The ViroBlast (Deng *et al.*, 2007) tool allows users to search for entries in ProtVirDB that have sequence similarity to query protein sequences. This provides the advantage of parsing the results according to an *E*-value or score cut-off chosen by the user. The integrated ClustalW (Thomson *et al.*, 1994) and Muscle (Edgar, 2004) tools, supplemented with the colorful display generated by Jalview (Clamp *et al*., 2004), perform multiple sequence alignment of selected sequences. The Java-based ATV program (Clamp *et al.*, 2004) allows the viewing of phylogenetic trees obtained from the QuickTree program (Howe *et al*., 2002). The antigenic program from the EMBOSS package (Rice *et al.*, 2000) predicts potentially antigenic regions of a protein sequence. The detection of conserved motifs in protein sequences is critical for annotation of proteins. The available tools for this purpose like PPsearch (http://www.ebi.ac.uk/Tools/ppsearch/index.html) allow the user to scrutinize only the already recognized and conserved motifs in databases like Prosite. Herein, a simple and versatile tool called ProbeMotiF facilitates search of user-defined motifs within the ProtVirDB database or any other user-defined set of sequences. This is a PERL-based tool that allows the users to search for motifs using regular expressions including wildcards. This serves as a supplementary tool, especially in cases where a newly discovered motif can be quickly searched in the database.

5 AMINO ACID REPEATS ANALYSIS

Amino acid repeats have been correlated with virulence in bacteria as well as protozoans (Fankhauser *et al*., 2007). We scanned the entire ProtVirDB database for the presence of repeats using the DIREP program developed by us earlier (Kalita *et al*., 2006). Interestingly, both homo- and heterorepeats were detected in 32% of the total sequences (101 unique proteins out of 315). See Supplementary Materials 3 and 4 for the list of sequences. In the database, the *P. falciparum* sequences alone accounted for 33 proteins with repeats (out of 45), followed by *Entamoeba histolytica* 11 (out of 29) (see Supplementary Material 1 for details). These figures are strikingly high when compared with the percentages of repeat containing proteins in the entire proteomes [33.49% and 2.79%, respectively, Depledge *et al*. (2007)]. This may well be an underestimation of the proportions since the set of ProtVirDB virulent proteins represents only the currently annotated virulent proteins in the parasite genomes. Yet, the presence of repeats in almost one-third of the proteins within a small collection is certainly intriguing and reinforces that repeat-containing proteins play an indispensable role in the parasite's virulence. It is noteworthy that we did not observe any bias of repeat-containing proteins within any specific functional category. Cysteine proteases, proteases and heat shock protein categories were scantily represented in this set, but this bias could be due to their under-representation in the database.

6 PERSPECTIVES

Diseases caused by parasitic protozoans are often studied in isolation; however, comparative studies may provide a key to hitherto undiscovered but common mechanisms of virulence. The ProtVirDB database can assist in research efforts aimed at such comparative studies. It also provides ground for further studies related to the significance of repeat-containing proteins in the virulence of the protozoan parasites. The database will be updated regularly and additional tools incorporated. Given the mounting interest in protozoan parasitic diseases, we expect ProtVirDB to serve as a valuable resource to the scientific community.

ACKNOWLEDGEMENTS

Computational resources were provided by the Bioinformatics Laboratory, Structural and Computational Biology Group, International Center for Genetic Engineering and Biotechnology, New Delhi, India.

Funding: Council of Scientific and Industrial Research (CSIR) fellowship (to J.R.).

Conflict of Interest: none declared.

REFERENCES

Aguero,F. *et al*. (2008) Genomic scale prioritization of drug targets: the TDR targets database. *Nat. Rev. Drug Discov.*, **7**, 900–907.

Clamp,M. *et al*. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.

Deng,W. *et al*. (2007) ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. *Bioinformatics*, **23**, 2334–2336.

Depledge,D.P. *et al*. (2007) RepSeq-a database of amino acid repeats present in eukaryotic pathogens. *BMC Bioinformatics*, **8**,122.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Fankhauser,N. *et al*. (2007) Surface antigens and potential virulence factors from parasites detected by comparative genomics of perfect amino acid repeats. *Proteome Sci.*, **5**, 20.

Finn,R.D. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

Howe,K. *et al*. (2002) QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics*, **18**, 1546–1547.

Kalita,M.K. *et al. (*2006) ProtRepeatsDB: a database of amino acid repeats in genomes. *BMC Bioinformatics*, **7**, 336.

Prescott,L.M. *et al*. (1999) Symbiotic associations: parasitism, pathogenicity, and resistance. Kane,K.T. and Smith,J.M. (eds) *Microbiology*, 4th edn. WCB McGraw-Hill, Boston, pp. 581–591.

Rice,P. *et al*. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Thompson,J.D. *et al*. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment though sequence weighting ,position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Zhou,C.E. *et al*. (2006) MVirDB – a microbial database of protein toxins,virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.*, **35**, D391–D394.