Published online 7 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D391–D394***
doi:10.1093/nar/gkl791

MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications

C. E. Zhou*, J. Smith, M. Lam, A. Zemla, M. D. Dyer[1] and T. Slezak

Pathogen Bioinformatics, Mailstop L-174, Energy, Environment, and Biology Division, Computations Directorate, Lawrence Livermore National Laboratory, 70000 East Avenue, Livermore, CA 94550, USA and [1]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Received June 5, 2006; Revised August 28, 2006; Accepted October 2, 2006

ABSTRACT

Knowledge of toxins, virulence factors and antibiotic resistance genes is essential for bio-defense applications aimed at identifying 'functional' signatures for characterizing emerging or engineered pathogens. Whereas genetic signatures identify a pathogen, functional signatures identify what a pathogen is capable of. To facilitate rapid identification of sequences and characterization of genes for signature discovery, we have collected all publicly available (as of this writing), organized sequences representing known toxins, virulence factors, and antibiotic resistance genes in one convenient database, which we believe will be of use to the bio-defense research community. MvirDB integrates DNA and protein sequence information from Tox-Prot, SCORPION, the PRINTS virulence factors, VFDB, TVFac, Islander, ARGO and a subset of VIDA. Entries in MvirDB are hyperlinked back to their original sources. A blast tool allows the user to blast against all DNA or protein sequences in MvirDB, and a browser tool allows the user to search the database to retrieve virulence factor descriptions, sequences, and classifications, and to download sequences of interest. MvirDB has an automated weekly update mechanism. Each protein sequence in MvirDB is annotated using our fully automated protein annotation system and is linked to that system's browser tool. MvirDB can be accessed at http://mvirdb.llnl.gov/.

INTRODUCTION

The anthrax bio-terror event in October 2001 brought into focus the need for reagents to rapidly identify pathogenic organisms. There is potential for bio-terror events involving genetically modified organisms with enhanced virulence or resistance to current antibiotic therapies. This emphasizes the urgent need for reagents to specifically identify factors that could be artificially introduced into a pathogen, or an otherwise benign organism, thus conferring enhanced pathogenicity. Therefore, in recent years there has been a surge of interest in the bio-defense community in identifying and characterizing toxins, virulence factors and antibiotic resistance genes in order to recognize 'functional' signatures for detecting emerging or engineered pathogens. Whereas genomic signatures identify a pathogen, functional signatures identify what a pathogen is capable of. MvirDB allows us to rapidly identify and evaluate sets of genes that are of highest priority to target in our design of DNA and protein signatures. In order to facilitate rapid identification and characterization of important sequences for signature discovery, we integrated all publicly available, organized sequences representing known toxins, virulence factors and antibiotic resistance genes into one convenient database.

As of this writing, there are eight public-access sequence databases comprising the known protein toxins, virulence factors and antibiotic resistance genes: the Tox-Prot (known protein toxins) subset of the Swiss-Prot protein knowledgebase (1), the SCORPION molecular database of scorpion toxins (2), the PRINTS database of virulence factors (3,4), the VFDB virulence factor database (5), the TVFac toxin and virulence factor database at Los Alamos National Laboratory, the Islander database of genomic islands (6), the ARGO database of vancomycin and b-lactam antibiotic resistance genes (7), and the VIDA database of animal virus open reading frames, comprising five virus families (8). MvirDB is a data warehouse integrating organism, keyword, classification, cross-reference, attribute and sequence from all of these databases. Because genetic engineering prompts the need for detection reagents specific for toxins from any taxonomic origin, we did not limit our efforts to bacterial toxins, but considered any protein toxin from any organism—hence our inclusion of SCORPION and non-bacterial toxins in Tox-Prot. In a similar vein, we adopted a broad definition of 'virulence' and reasoned that proteins that mediate

*To whom correspondence should be addressed. Tel: +1 925 422 2117; Fax: +1 925 423 6437; Email: zhou4@llnl.gov pathogen–host interaction, such as viral proteins that may interact with human cell surface receptors, are of interest. Thus, we extracted a subset of VIDA genes consisting of the host–virus interaction genes for the Herpesviridae, Papillomaviridae and Poxviridae virus families. Because virulence genes are often contained within genomic islands, we also included the Islander data set. New data sources planned for inclusion in MvirDB will be posted on the web site.

Our motivations in creating MvirDB were to (i) centralize sequence information of interest in bio-defense, (ii) facilitate retrieval, which would otherwise require navigation to several web sites and use of various tools to extract individual entries, (iii) provide a comprehensive set of sequences against which we could automate the application of sequence comparison tools (e.g. Blast), (iv) facilitate analyses using our high-throughput annotation system (http://manndb.llnl.gov/), (v) enable structural studies and classification and (vi) input and categorize proprietary sequences representing virulence genes provided by collaborators. We have been using this database for several years in our own bio-defense applications (9,10), and we believe it will be useful to the general bio-defense research community as well as to researchers in medicine. Here we present the construction of MvirDB and describe the tools that enable navigation and analysis.

Construction and content

MvirDB is implemented as an Oracle 10g relational database. The schema for MvirDB data organization is available on our website: http://mvirdb.llnl.gov/. Figure 1 is a dataflow diagram for construction of MvirDB illustrating data acquisition, parsing, loading, analysis and viewing. A synchronous process, running weekly, queries each of the eight public-access databases and pulls data in hypertext format. Data of interest are then recognized, copied out by parsers that are specific for each data source, and loaded into MvirDB; selected data fields include: gene name, short and long descriptions, organization (e.g. hierarchical organization of PRINTS virulence factors), keywords, DNA and/or protein sequence, links back to original data sources, links (when available) to other external databases (e.g. PIR, COGS). Each entry is tagged (given a 'status' designation) to indicate whether it is a known toxin, virulence factor or antibiotic resistance gene (e.g. entries from VFDB), or whether its status is unknown (e.g. VIDA entries). We do not coalesce entries when data redundancy resulting from dual classification in an underlying data source is identified or when identical sequences (or essentially identical entries) occur in two or more underlying data sources, reasoning that each 'redundant' entry has been curated (i.e. classified or annotated) differently. Hierarchical classifications of entries in the PRINTS database are preserved in MvirDB, along with associated sub-sequences (fingerprints) for each entry. Entries that include unique Genbank identifiers are compared to all entries in our microbial annotation database (MannDB); a foreign key link is established in MvirDB to any MannDB entry that has the same Genbank identifier. All protein sequences in MvirDB are taxonomically grouped in order to facilitate batch processing through MannDB annotation tools; each grouping is entered into MannDB, and each protein sequence is fully annotated using our high-throughput annotation system. Taxonomic grouping is necessary in order to correctly specify parameters for input to certain annotation tools that modify the analysis based on which taxon is represented or for running tools that are specific for certain taxa. Hyperlinks to the MannDB browser are established to enable convenient viewing of annotation results. Using a synchronous process, running weekly, protein and DNA entries are separately saved to a Unix file and formatted as Blast databases using formatdb (11). A web-based client browser (Figure 2) allows the user to search the contents of MvirDB and view data, and a web-based client Blast interface allows the user to blast a DNA or protein sequence against MvirDB. MvirDB currently holds 9059 entries from 1220 organisms.

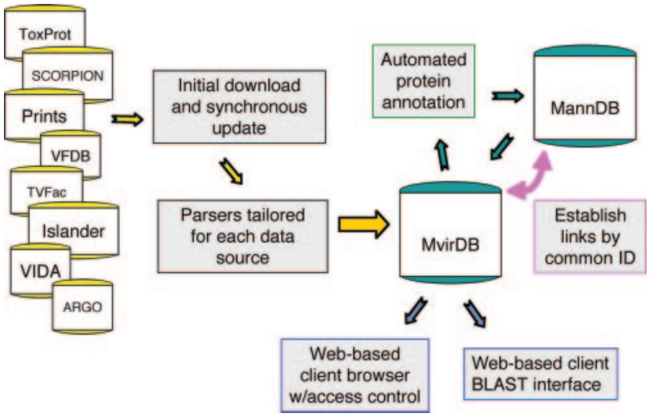
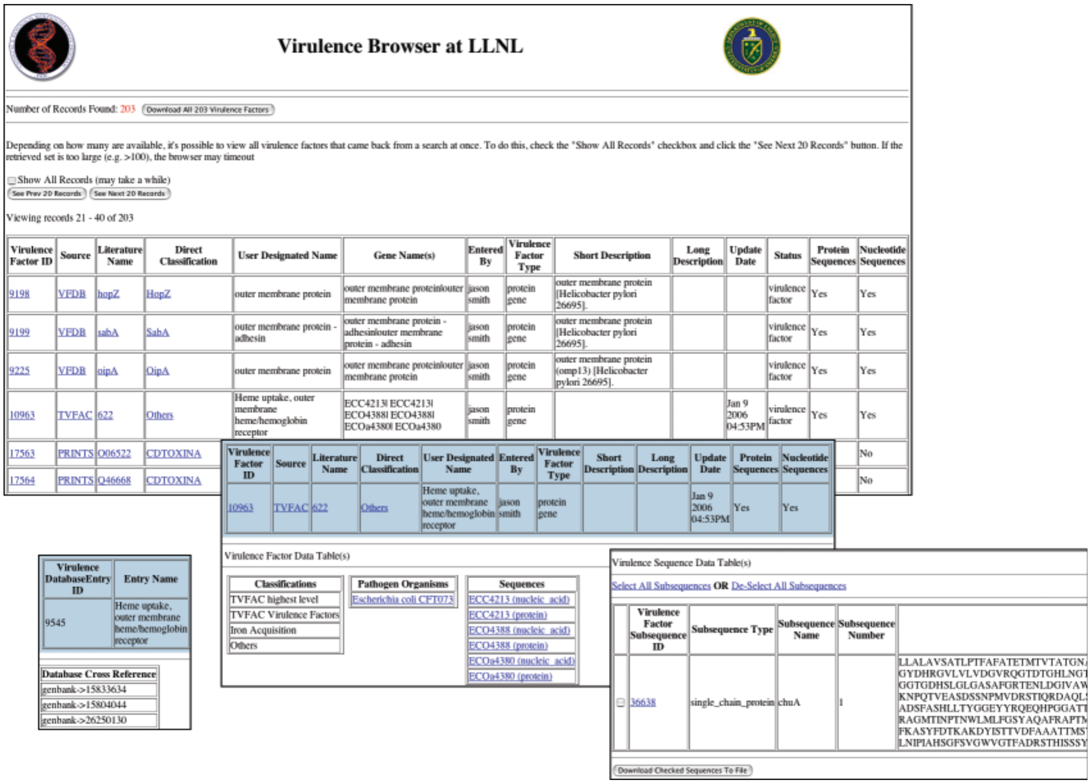

ToxProt
SCORPION
Prints
VFDB
TVFac
Islander
VIDA
ARGO
Initial download and synchronous update
Parsers tailored for each data source
Automated protein annotation
MannDB
MvirDB
Establish links by common ID
Web-based client browser w/access control
Web-based client BLAST interface

Figure 1. Dataflow diagram for MvirDB. Public-access and proprietary data sources are downloaded and parsed into MvirDB. Protein entries in MvirDB are sorted by organism type (e.g. Gram-positive bacterium, eukaryotic virus) and annotated using the MannDB automated annotation system. Additional links are established between MvirDB and MannDB when entries contain common external database identifiers. Web client browser and BLAST interfaces enable viewing and analysis of data.

Utility and discussion

MvirDB's blast interface allows the user to search for entries in MvirDB that have sequence similarity to a DNA or protein sequence of interest. The user may specify Blast parameters and upload or paste the query sequences. Results are returned in a readable table format and ordered by ascending *E*-value. Each hit from MvirDB is hyperlinked to that entry's browser page. The Blast formatted MvirDB database is used for automated identification of similar sequence entries in MannDB. The MvirDB query page provides several fields that the user can select for constructing queries against MvirDB: virulence factor ID or name, taxonomic group, organism, classification, keyword, synonym, or status. The user can then use the browser to locate information about the entry, link to MannDB annotations, link to additional information on external websites, or download sequences to a file.

MvirDB can be used for identifying and characterizing sequences of interest in bio-defense applications, such as design of reagents for threat agent detection (9,12). The integration of data from eight databases (each with its own format and toolset) greatly simplifies searches for



Virulence Browser at LLNL

Number of Records Found: 203 (Download All 203 Virulence Factors)

Depending on how many are available, it's possible to view all virulence factors that came back from a search at once. To do this, check the "Show All Records" checkbox and click the "See Next 20 Records" button. If the retrieved set is too large (e.g. >100), the browser may timeout

□ Show All Records (may take a while)

(See Prev 20 Records) (See Next 20 Records)

Viewing records 21 - 40 of 203

Virulence Factor ID	Source	Literature Name	Direct Classification	User Designated Name	Gene Name(s)	Entered By	Virulence Factor Type	Short Description	Long Description	Update Date	Status	Protein Sequences	Nucleotide Sequences
9198	VFDB	hopZ	HopZ	outer membrane protein	outer membrane proteinlouter membrane protein	jason smith	protein gene	outer membrane protein [Helicobacter pylori 26695].			virulence factor	Yes	Yes
9199	VFDB	sabA	SabA	outer membrane protein - adhesin	outer membrane protein - adhesinlouter membrane protein - adhesin	jason smith	protein gene	outer membrane protein [Helicobacter pylori 26695].			virulence factor	Yes	Yes
9225	VFDB	oipA	OipA	outer membrane protein	outer membrane proteinlouter membrane protein	jason smith	protein gene	outer membrane protein (omp13) [Helicobacter pylori 26695].			virulence factor	Yes	Yes
10963	TVFAC	622	Others	Heme uptake, outer membrane heme/hemoglobin receptor	ECC4213l ECC4213l ECO4388l ECO4388l ECOa4380l ECOa4380	jason smith	protein gene			Jan 9 2006 04:53PM	virulence factor	Yes	Yes
17563	PRINTS	O06522	CDTOXINA										No
17564	PRINTS	Q46668	CDTOXINA										No

Virulence Factor ID	Source	Literature Name	Direct Classification	User Designated Name	Entered By	Virulence Factor Type	Short Description	Long Description	Update Date	Protein Sequences	Nucleotide Sequences
10963	TVFAC	622	Others	Heme uptake, outer membrane heme/hemoglobin receptor	jason smith	protein gene			Jan 9 2006 04:53PM	Yes	Yes

Virulence Factor Data Table(s)

Classifications
TVFAC highest level
TVFAC Virulence Factors
Iron Acquisition
Others

Pathogen Organisms
Escherichia coli CFT073

Sequences
ECC4213 (nucleic_acid)
ECC4213 (protein)
ECO4388 (nucleic_acid)
ECO4388 (protein)
ECOa4380 (nucleic_acid)
ECOa4380 (protein)

Virulence DatabaseEntry ID	Entry Name
9545	Heme uptake, outer membrane heme/hemoglobin receptor

Database Cross Reference
genbank->15833634
genbank->15804044
genbank->26250130

Virulence Sequence Data Table(s)

Select All Subsequences **OR** De-Select All Subsequences

	Virulence Factor Subsequence ID	Subsequence Type	Subsequence Name	Subsequence Number	
□	36638	single_chain_protein	chuA	1	LLALAVSATLPTFAFATETMTVTATGNA GYDHRGVLVLVDGVRQGTDTGHLNGT GGTGDHSLGLGASAFGRTENLDGIVAW KNPQTVEASDSSNPMVDRSTIQRDAQLS ADSFASHLLTYGGEYYRQEQHPGGATT RAGMTINPTNWLMLFGSYAQAFRAPTM FKASYFDTKAKDYISTTVDFAAATTMS LNIPIAHSGFSVGWVGTFADRSTHISSSY

(Download Checked Sequences To File)

Figure 2. Virulence database browser sample web pages. An example free-text query on 'Outer membrane' yields 203 entries from VFDB, TVFAC and PRINTS data sources. User can then select a virulence factor (e.g. 10 963; center panel) and display a sequence (lower right panel). Linking from the sequence display page via the virulence database entry ID displays external database cross-reference IDs (lower left panel).

sequence information relating to toxins, virulence factors, and antibiotic resistance genes, and enables high-throughput annotation, which is of value in characterizing proteins as potential targets for directing design of reagents for detection or drugs for therapeutics (10). MvirDB does not attempt to replicate or supercede the databases it taps, but rather collects enough information to assist the user in identifying and evaluating a comprehensive set of sequences using a single browser interface. Having used MvirDB for several years in bio-defense applications, we believe it will serve as a valuable resource for others in the bio-defense and medical research communities.

CONCLUDING REMARKS

MvirDB is a centralized resource (data warehouse) comprising all publicly accessible, organized sequence data for protein toxins, virulence factors and antibiotic resistance genes. Protein entries in MvirDB are annotated using a high-throughput, fully automated computational annotation system; annotations are updated periodically to ensure that results are derived using current public database and open-source tool releases. Tools provided for using MvirDB include a web-based browser tool and blast interface. We believe that MvirDB will be useful and convenient resource for researchers in the bio-defense and medical fields for rapidly acquiring information about protein toxin, virulence factor and antibiotic resistance gene sequences.

ACKNOWLEDGEMENTS

This work was performed under the auspices of the U.S. Department of Energy by the University of California Lawrence Livermore National Laboratory under contract no. W-7405-ENG-48 and was supported by funding from the U.S. Department of Homeland Security. Funding to pay the Open Access publication charges for this article was provided by U.S. Department of Homeland Security.

Conflict of interest statement. None declared.

REFERENCES

1. Jungo,F. and Bairoch,A. (2005) Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase. *Toxicon*, **45**, 293–301.
2. Srinivasan,K.N., Gopalakrishnakone,P., Tan,P.T., Chew,K.C., Cheng,B., Kini,R.M., Koh,J.L.Y., Seah,S.H. and Brusic,V. (2001) SCORPION, a molecular database of scorpion toxins. *Toxicon*, **40**, 23–31.
3. Paine,K. and Flower,D.R. (2002) Bacterial bioinformatics: pathogenesis and the genome. *J. Mol. Biotechnol.*, **4**, 357–365.
4. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement, preprints. *Nucleic Acids Res.*, **31**, 400–402.
5. Chen,L., Yang,J., Yu,J., Yao,Z., Sun,L., Shen,Y. and Jin,Q. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, **33**, D325–D328.

6. Mantri,Y. and Williams,K.P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.*, **32**, D55–D58.
7. Scaria,J., Ghandramouli,U. and Verma,S.K. (2005) Antibiotic resistance genes online (ARGO): a database on vancomycin and b-lactam resistance genes. *Bioinformation*, **1**, 5–7.
8. 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.
9. Slezak,T., Kuczmarski,T., Ott,L., Torres,C., Mederos,D., Smith,J., Truitt,B., Mulakken,N., Lam,M., Vitalis,E. *et al.* (2003) Comparative genomics tools applied to bioterrorism defense. *Brief. Bioinform.*, **4**, 133–149.
10. Zhou,C.L.E., Zemla,A., Roe,D., Young,M., Lam,M., Schoeniger,J.S. and Balhorn,R. (2005) Computational approaches for identification of conserved/unique binding pockets in the A chain of ricin. *Bioinformatics*, **21**, 3085–3096.
11. Altschul,S.F., Gish,W., Miller,W., Meyers,E.W. and Lipman,D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **5**, 403–410.
12. Fitch,J.P., Gardner,S.N., Kuczmarski,T.A., Kurtz,S., Myers,R., Ott,L.L., Slezak,T.R., Vitalis,E.A., Zemla,A.T. and McCready,P.M. (2002) Rapid development of nucleic acid diagnostics. *Proc. IEEE*, **90**, 1708–1720.