BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 21 2008, pages 2564–2565
doi:10.1093/bioinformatics/btn477

Databases and ontologies

varDB: a pathogen-specific sequence database of protein families involved in antigenic variation

C. Nelson Hayes[1], Diego Diez[1], Nicolas Joannin[2,3], Wataru Honda[1], Minoru Kanehisa[1], Mats Wahlgren[2,3], Craig E. Wheelock[1,2,4,*] and Susumu Goto[1,*]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan, [2]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Box 280, SE-17177 Stockholm, [3]Swedish Institute for Infectious Disease Control (SMI), SE-17182 Stockholm and [4]Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-17177 Stockholm, Sweden

Received on July 18, 2008; revised on August 28, 2008; accepted on September 4, 2008

Advance Access publication September 6, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Infectious diseases are a major threat to global public health and prosperity. The causative agents consist of a suite of pathogens, ranging from bacteria to viruses, including fungi, helminthes and protozoa. Although these organisms are extremely varied in their biological structure and interactions with the host, they share similar methods of evading the host immune system. Antigenic variation and drift are mechanisms by which pathogens change their exposed epitopes while maintaining protein function. Accordingly, these traits enable pathogens to establish chronic infections in the host. The varDB database was developed to serve as a central repository of protein and nucleotide sequences as well as associated features (e.g. field isolate data, clinical parameters, etc.) involved in antigenic variation. The data currently contained in varDB were mined from GenBank as well as multiple specialized data repositories (e.g. PlasmoDB, GiardiaDB). Family members and ortholog groups were identified using a hierarchical search strategy, including literature/author-based searches and HMM profiles. Included in the current release are >29 000 sequences from 39 gene families from 25 different pathogens. This resource will enable researchers to compare antigenic variation within and across taxa with the goal of identifying common mechanisms of pathogenicity to assist in the fight against a range of devastating diseases.

Availability: varDB is freely accessible at http://www.vardb.org/

Contact: nelson@kuicr.kyoto-u.ac.jp; goto@kuicr.kyoto-u.ac.jp

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

1 INTRODUCTION

Infectious diseases including AIDS, tuberculosis and malaria are collectively responsible for significant mortality and morbidity on a global scale. While the individual pathogens come from a range of divergent taxa, they share the common mechanisms of antigenic variation and drift to avoid clearance by the host immune system. These mechanisms are ruled by two evolutionary pressures: maintaining protein function and varying the antigens present within the proteins. For example, influenza undergoes antigenic drift and therefore requires continual adjustment to the vaccine formulation (Oxford *et al*., 2003), whereas *Plasmodium* uses surface antigen switching to maintain a pool of surface receptors that vary sufficiently to avoid the development of cross-reactive antibodies (Kaviratne *et al*., 2003). The large size of these protein families, combined with the rapid evolution of the variant genes within strains, leads to an immense collection of unique antigens.

varDB was developed to serve as a centralized database of gene families involved in antigenic variation. While there are a number of pathogen- or disease-specific sequence databases, e.g. PlasmoDB (Bahl *et al*., 2003) and the Los Alamos HIV Database (Hraber *et al*., 2007), to our knowledge there is no database specifically dedicated to antigenic variation that encompasses multiple taxonomic groups. However, common strategies for generating phenotypic diversity in response to immune defenses have been noted among diverse pathogens (Deitsch *et al*., 1997). Consequently, there are a number of potential advantages to having a common repository of antigenic variable sequences including: (i) the ability to compare common mechanisms across diverse taxa, (ii) centralized access to an expanding repertoire of isolate and genome project data and (iii) the development of specialized tools for the analysis of hyper-variable nucleic acid and protein sequences.

2 DATA COLLECTION

Pathogens of medical and veterinary importance demonstrating antigenic variation were selected for inclusion in varDB. Table 1 shows several representative taxa included in the database, and additional pathogens will be added with continued development.

The identification of antigenic variant gene families was based on published reports in the scientific literature. Sequences and annotations were downloaded from DDBJ/EMBL/GenBank based on taxon-specific identifiers and separated either as isolate-specific or genome project sequences. Isolate sequences were collected based on sequence similarity profiles, and keyword searches based on author, publication and gene/protein annotations. Genome sequences were searched for similarity to the target gene

*To whom correspondence should be addressed.

Table 1. Selected organisms and antigenic variation gene families in varDB

Taxonomic group	Species	Disease	Gene families	N° sequences
Bacteria	*Anaplasma spp.*	Anaplasmosis	*msp2*	141
	Ehrlichia spp.	Canine ehrlichiosis	*msp4*	94
Fungi	*Pneumocystis carinii*	Pneumonia	*msg*	73
Helminthes	*Echinococcus granulosus*	Cystic echinococcosis	*antigen B*	153
Protozoa	*Giardia lambia*	Giardiasis	*vsp*	384
	Plasmodium spp.	Malaria	*var, rifin/stevor, pir*	23,719
	Trypanosoma brucei	African trypanosomiasis	*vsg*	221
Viruses	*HIV-1*	HIV/AIDS	*env, gag, nef*	3466

families using HMM models. GenBank records for some genome sequencing projects may be out of date due to progressive re-annotation following initial submission. In this case, current data were retrieved directly from project or species-specific centers (e.g. the Broad Institute Microbial Sequencing Center). A detailed description of the process is available in Figure S1. varDB will be updated on a regular basis as new genome sequences and field isolates become available.

3 DATA ANALYSIS

The main functionalities of varDB include querying and retrieving antigenic variation sequence data for comparative analysis. To this end, a keyword search and local BLAST database permit querying, sorting and filtering of sequences using a variety of criteria, e.g. species, strain, chromosome, Pfam domain, etc. (Figure S2). Filtering by collection date or geographic region is also possible and may provide insight into the degree of sequence variation within and among populations as well as over time.

Using the integrated shopping cart tools, sequences can be selected and organized into subsets and aligned using MAFFT (Katoh *et al*., 2005). Precomputed protein and DNA alignments can be downloaded or viewed online using either Jalview (Clamp *et al*., 2004) or a browser-based alignment viewer. Sequences can be downloaded in several standard formats, and sequence annotations can be downloaded in a tab-delimited property file.

4 SYSTEM IMPLEMENTATION

Sequences and annotation data are stored in a PostgreSQL (http://www.postgresql.org/) database. varDB is deployed on a JBoss application server (http://www.jboss.org) running under Linux. The Ajax-driven web interface is based on open source tools including BioJava (http://biojava.org/), the Spring Framework (http://www.springframework.org/) and the Ext JS library (http://extjs.com/).

5 PERSPECTIVES

Infectious diseases and their pathogens are often studied in isolation, but observations made in one species may yield insights in other taxa. The ability to perform comparative analyses over diverse taxa should assist in elucidating biological mechanisms of pathogenicity and increase our ability to identify new biological pathways, drug targets and therapeutic interventions. This version of varDB provides a framework for retrieving and storing sequences from different gene families from diverse species and multiple data sources. Building upon this foundation, varDB will expand the breadth and depth of the sequence coverage as well as specific tools for the analysis of antigenic variation. This new resource should assist in efforts to understand the fundamental biological mechanisms behind the observed pathogenicities as well as increase our ability to combat these devastating diseases.

ACKNOWLEDGEMENTS

Computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Funding: Vinnova-SSF Multidisciplinary BIO; 21st Century COE Program and Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan; STINT Foundation; PREGVAX, FP7-Health-2007-A-201588; Swedish Royal Academy of Sciences; Japanese Society for the Promotion of Science post-doctoral fellowships to C.N.H. and D.D.

Conflict of Interest: none declared.

REFERENCES

Bahl,A. *et al*. (2003) PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.

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

Deitsch,K.W. *et al*. (1997) Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. *Microbiol. Mol. Biol. Rev.*, **61**, 281–293.

Hraber,P.T. *et al*. (2007) Los Alamos hepatitis C virus sequence and human immunology databases: an expanding resource for antiviral research. *Antivir. Chem. Chemother.*, **18**, 113–123.

Katoh,K. *et al*. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

Kaviratne,M. *et al*. (2003) *Antigenic Variation in Plasmodium falciparum and Other Plasmodium Species*. Academic Press, Amsterdam/Boston.

Oxford,J. *et al*. (2003) *Influenza – the Chameleon Virus*. Academic Press, Amsterdam/Boston.