Published online 19 October 2010 Nucleic Acids Research, 2011, Vol. 39, Database issue D1073–D1078
doi:10.1093/nar/gkq944

Protegen: a web-based protective antigen database and analysis system

Brian Yang[1,2,3], Samantha Sayers[1,2,4], Zuoshuang Xiang[1,2,5] and Yongqun He[1,2,5,*]

[1]Unit for Laboratory Animal Medicine, University of Michigan Medical School, [2]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, [3]Harvard College, Harvard University, Cambridge, MA 02138, [4]College of Literature, Science, and the Arts, University of Michigan and [5]Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI 48109, USA

Received August 16, 2010; Revised September 16, 2010; Accepted September 29, 2010

ABSTRACT

Protective antigens are specifically targeted by the acquired immune response of the host and are able to induce protection in the host against infectious and non-infectious diseases. Protective antigens play important roles in vaccine development, as biological markers for disease diagnosis, and for analysis of fundamental host immunity against diseases. Protegen is a web-based central database and analysis system that curates, stores and analyzes protective antigens. Basic antigen information and experimental evidence are curated from peer-reviewed articles. More detailed gene/protein information (e.g. DNA and protein sequences, and COG classification) are automatically extracted from existing databases using internally developed scripts. Bioinformatics programs are also applied to compute different antigen features, such as protein weight and pI, and subcellular localizations of bacterial proteins. Presently, 590 protective antigens have been curated against over 100 infectious diseases caused by pathogens and non-infectious diseases (including cancers and allergies). A user-friendly web query and visualization interface is developed for interactive protective antigen search. A customized BLAST sequence similarity search is also developed for analysis of new sequences provided by the users. To support data exchange, the information of protective antigens is stored in the Vaccine Ontology (VO) in OWL format and can also be exported to FASTA and Excel files. Protegen is publically available at http://www.violinet.org/protegen.

INTRODUCTION

Human and animal health is threatened by various diseases every day. Acquired as a result of pathogenic microbial agents, infectious diseases are still a major source of mortality throughout the world, contributing to 26% of global mortality in 2001 (1). Of these mortalities, 90% are caused by illnesses such as acute respiratory infections, diarrheal diseases, malaria, AIDS, tuberculosis and measles. Cancer, allergy and many other diseases also cause significant mortality and morbidity in human and animal victims.

Vaccines stimulate the immune system and confer protection against pathogenic microorganisms, offering a safe effective method to prevent disease. Vaccines are among the most useful and cost-effective tools for reducing the morbidity and mortality caused by infectious diseases. Vaccination programs have succeeded in eliminating smallpox from the world population as well as in drastically reducing the incidences of other diseases such as polio. Measles, one of the six major illnesses responsible for mortality mentioned above, can now be prevented by vaccination. Vaccines are also being developed for fighting against other diseases such as cancer and allergy.

At the forefront of this vaccine development are protective antigens. Protective antigens are those antigens that are specifically targeted by the acquired immune response of the host, and when introduced into the host body, are able to stimulate the production of antibodies and/or cell-mediated immunity against certain pathogens or the causes of other diseases. Protective antigens can be used in many research areas. First, identification of protective antigens is a major component of research for new and improved vaccines against infectious diseases (2). Using these protective antigens, researchers are able to develop vaccines, such as DNA and subunit vaccines, which use targeted antigen DNA or proteins to elicit a

*To whom correspondence should be addressed. Tel: +1 734 615 8231; Fax: +1 734 936 3235; Email: yongqunh@umich.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

protective immune response. These vaccines have shown great potential in the fight against diseases such as malaria for which there is no current vaccine. Protective antigens are also useful in the control of allergies through the use of allergen immunotherapy. This process uses repeated exposure to allergenic antigens in order to create immunologic tolerance in an individual to the allergen (3). Many cancer vaccines are currently in clinical trials (4). In addition, protective antigens can often be used as biological markers for diagnosis of diseases such as AIDS based on the presence or absence of a protective antigen(s) (5). Furthermore, the identification of protective antigens and the study of their roles in the induction of host immunity against various diseases are crucial to the elucidation of fundamental host immune mechanisms.

While intensive research has been conducted and resulted in identification of many protective antigens, these is no reported central resource that allows storage, annotation, comparison, and analysis of protective antigens across different species. To address this challenge, we have developed Protegen (http://www.violinet.org/protegen), a web-based protective antigen database and analysis system. Protegen stores manually curated protective antigens and associated information. It also allows bioinformatics analysis and comparison of various protective antigens.

SYSTEM AND DATABASE DESIGN

Protegen is implemented using a three-tier architecture built on two Dell Poweredge 2580 servers which run the Redhat Linux operating system (Redhat Enterprise Linux ES 4). Users can submit database or analysis queries through the web. These queries are then processed using PHP/SQL (middle-tier, application server based on Apache) against a MySQL (version 5.0) relational database (back-end, database server). The result of each query is then presented to the user in the web browser. Two servers are scheduled to regularly backup each others' data. Protegen is an integrated program of the VIOLIN vaccine database and analysis (6).

Supplementary Figure S1 demonstrates the Protegen database schema. For each specific protective antigen, the Protegen database contains the following information: (i) Detailed reference citation information, for example, authors, journal, title, year, volume, issue, and pages. These references are typically obtained from PubMed (http://www.ncbi.nlm.nih.gov/pubmed), and used to obtain experimental evidence that confirms a protein's molecular role as a protective antigen. (ii) General information of protective antigens, for example, gene symbol, protein name, protein function, NCBI Gene ID, NCBI Nucleotide ID, NCBI Protein GI, locus tag, NCBI taxonomy ID, Cluster of Orthologous Groups (COG) category (7), chromosome number, segment number, plasmid number, gene starting and ending position, gene strand orientation and DNA and protein sequences. These data can be automatically extracted and updated from existing databases (e.g. NCBI RefSeq database) using an internally developed program. (iii) Protein 3D structure stored in the Protein Data Bank (http://www.pdb.org). This information is manually curated. (iv) Calculated results (e.g. protein weight and pI, and subcellular localizations of bacterial protective antigens) based on DNA or protein sequences with reliable bioinformatics software programs. (v) Customized BLAST libraries containing DNA or protein sequences of all protective antigens in Protegen. The customized Protegen BLAST libraries can be used for BLAST sequence similarity searches.

SEMI-AUTOMATIC ANNOTATION OF PROTECTIVE ANTIGENS

A semi-automatic annotation system is developed in Protegen for protective antigen curation and analysis (Figure 1). This annotation system includes manual curation from selected peer-reviewed publications and bioinformatics extraction and analysis of DNA and protein sequences of protective antigens. Manual curation of peer-reviewed journal articles emphasizes the retrieval of antigen information [e.g. protein/gene names and database identifiers (IDs)] and experimental evidence. Strong experimental evidence, particularly results from a protection assay against specific challenge or an immune response assay that correlates with protection, is required to justify a protein being a protective antigen. It is a major task to identify and record such direct experimental evidence for individual protective antigens stored in Protegen. In many cases, curated journal articles contain database IDs for individual proteins or genes, such as NCBI Gene ID, NCBI GenBank nucleotide ID or NCBI protein accession number. If no database ID is provided in curated publication, an ID can be found by search a NCBI database using the species/strain and protein/gene information from the article. Using a NCBI Gene/Nucleotide/Protein identifier, three scripts were developed to automatically extract related protein/gene information from corresponding database(s), such as DNA and protein sequence, gene description, and protein notes. These scripts can speed up the curation process and ensure data accuracy. The extracted results are automatically displayed on an online form and can be subject to further reviews and updates.

The DNA and protein sequences can further be used for more bioinformatics analyses (Figure 1). For example, protein weight and protein pI values are calculated using functions in BioPHP (http://genephp.sourceforge.net/). Subcellular localizations are predicted using PSORTb (8). The transmembrane helix domain analysis is performed using optimized HMMTOP (9). Adhesin probability is predicted using optimized SPAAN (10). The last three programs are also used in the Vaxign reverse vaccinology program (http://www.violinet.org/vaxign) developed by our group (11). Reverse vaccinology is an emerging vaccine design strategy that initiates the process by bioinformatically screening vaccine candidates from pathogenic genome databases based on different criteria (12). Vaxign is the first web-based system for genome-wide reverse vaccinology analysis. Subcellular localization, transmembrane domain and adhesin

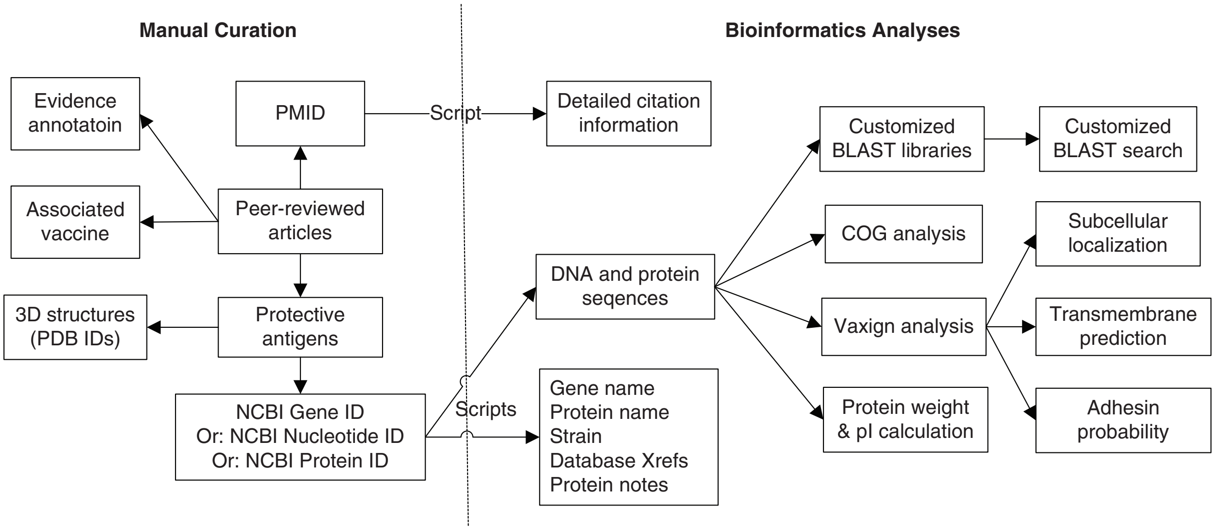


Figure 1. Semi-automatic annotation of protective antigens in Protegen overall design and architecture. Manual curation includes peer-reviewed publications from PubMed. A PubMed ID (PMID) is extracted and used to retrieve detailed citation information (e.g. authors, journal, and date). The evidence that proves the status of protective antigen for each protein is curated from published experimental studies. Vaccines associated with the protective antigens are also curated. PDB IDs are manually retrieved when available to provide 3D structure information of individual protective antigens. Internally developed script uses an input sequence ID from a NCBI database (e.g. NCBI Entrez Gene database) to automatically retrieve different types of information. The extracted DNA and protein sequences are further used for bioinformatics analyses using different methods.

probability are three important criteria for bacterial vaccine candidate prediction in reverse vaccinology. These programs are used in Protegen to support further comparison and analyses of known protective antigens.

The semi-automatic Protegen annotation system is developed by modifying an in-house web-based literature mining and curation system called Limix (13,14). The interactive Limix data submission and review system: (i) allows a curator to search literature, copy and edit text, and submit data to database and (ii) provides a data reviewer tools to review, edit, and approve the curated data on one comprehensive web interface. Limix also features automated reference tracking and management using a version control mechanism similar to that used in Wikipedia (http://www.wikipedia.org/). The system stores all the history changes made to every antigen. A reviewer can compare any two of the history versions to trace the detailed changes made to an antigen. Supplementary Figure S2 provides a more detailed view of the Limix curation process and reviewing features. Upon approval after the critical review, the data will be posted publicly.

Currently, Protegen has included 590 protective antigens. Among these protective antigens, 539 come from 44 bacteria, 40 viruses, 19 parasites and one fungal species, which cause various infectious diseases (Table 1 and Supplementary Table S1). Examples of pathogens include *Bacillus anthracis* (anthrax), *Brucella* spp. (brucellosis), *Yersinia pestis* (Plague), Human immunodeficiency virus (HIV, the causative agent of AIDS), Ebola virus (hemorrhagic fever) and *Plasmodium* spp. (malaria). Protegen also stores 59 protective antigens for cancer, allergies and a plant (ricin) toxin. Research into protective antigens against cancer and common allergens is an important field of research as both cancer and allergies cause mortality and morbidity just as infectious diseases do. Cancer vaccine development strategies focus on antigens to immunize cancer patients with to stimulate their own immune system, and thus are more treatment oriented than prevention based. The same approach is true for allergies, with the goal being to prevent an allergic response, in patients with known allergies to specific allergens.

Protegen is targeted to be updated quarterly with additional protective antigens and their bioinformatics analyses. All Protegen updates are posted in the Protegen website.

PROTECTIVE ANTIGEN DATA QUERY AND DISPLAY

The manually curated and pre-computed Protegen data can be efficiently queried and visualized as demonstrated in Figure 2. The protective antigens can be queried by specifying one or multiple criteria: (i) gene or protein name, (ii) pathogen species or strain, (iii) locus tag of a protective antigen, (iv) other existing database identifiers, such as NCBI Gene ID, NCBI nucleotide or protein GI, (v) keywords search with Boolean mode support, (vi) COG category, (vii) subcellular localization, (viii) four limiting factors, including the maximum number of transmembrane helices, the minimum adhesin probability, and protein sequence similarity to human proteins, and protein sequence similarity to mouse proteins (Figure 2A).

All query hits are first displayed on a web table containing basic antigen information (Figure 2B). Individual protective antigens can further be selected by a user to show more detailed information (Figure 2C). Other dynamic analysis of protective antigens, such as BLAST sequence similarity analysis, can be implemented (Figure 2D).

Table 1. Curated protective antigens from select pathogens as of 14 August 2010

Pathogen (Disease name)	Number of protective antigens
Bacteria (10 out of 44)	
Brucella spp. (Brucellosis)	19
B. pertussis (Whooping cough)	10
C. abortus (Pelvic inflammatory disease)	10
E. coli (Hemorrhagic colitis)	17
F. tularensis (Tularemia)	10
H. influenzae (Pneumonia, bacterial meningitis)	14
M. tuberculosis (Tuberculosis)	24
N. meningitidis (Meningitis)	19
S. pneumoniae (Pneumonia)	12
Y. pestis (Plague)	24
Viruses (10 out of 40)	
Dengue virus (Dengue fever)	4
Ebola virus (Hemorrhagic fever)	13
Herpes simplex virus type 1 and type 2 (Herpes)	9
Human Immunodeficiency Virus (AIDS)	5
Human papillomavirus (HPV)	4
Influenza virus (Influenza)	37
Japanese encephalitis virus (Japanese encephalitis)	6
Marburg virus (Hemorrhagic fever)	6
Pseudorabies virus (Aujeszky's disease)	8
Rotavirus (Severe diarrhea)	8
Parasites (10 out of 19)	
Babesia bovis (Babesiosis)	3
Eimeria tenella (hemorrhagic cecal coccidiosis)	5
Entamoeba histolytica (Amebiasis)	4
Leishmania amazonensis (Leishmaniasis)	5
Leishmania donovani (Visceral leishmaniasis)	11
Leishmania major (Cutaneous leishmaniasis)	10
Neospora caninum (Neosporosis)	8
Plasmodium spp. (Malaria)	26
Toxoplasma gondii (Toxoplasmosis)	3
Trypanosoma cruzi (Chagas Disease)	9
Others	
Allergy	14
Cancer	36
Ricin Toxin	1
Coccidioides spp. (Coccidioidomycosis)	9
Total	403

BLAST ANALYSIS OF PROTECTIVE ANTIGENS

To facilitate antigen comparison and dynamical sequence analysis, two customized BLAST libraries have been generated. One library contains the protein sequences of all protective antigens in Protegen, and the other collects all the DNA sequences of all protective antigens. The Protegen users are able to perform BLAST sequence similarity searches using different NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi), including blastp, blastn, blastx, tblastn and tblastx, against the customized Protegen libraries. The customized BLAST programs in Protegen can be used in two different applications. The first is to allow a user to compare a known protective antigen(s) in Protegen with other protective antigens. An example of such a BLAST search is demonstrated in Figure 2D. The second application is to allow a user to input any protein or DNA sequence(s) into the Protegen BLAST online form, and identify those protective antigens that share similar protein or DNA sequences with the input sequence(s).

Other dynamic analyses of protective antigens are also available. For example, each protective antigen can be linked to Vaxign for more reverse vaccinology analyses (e.g. immune epitope prediction).

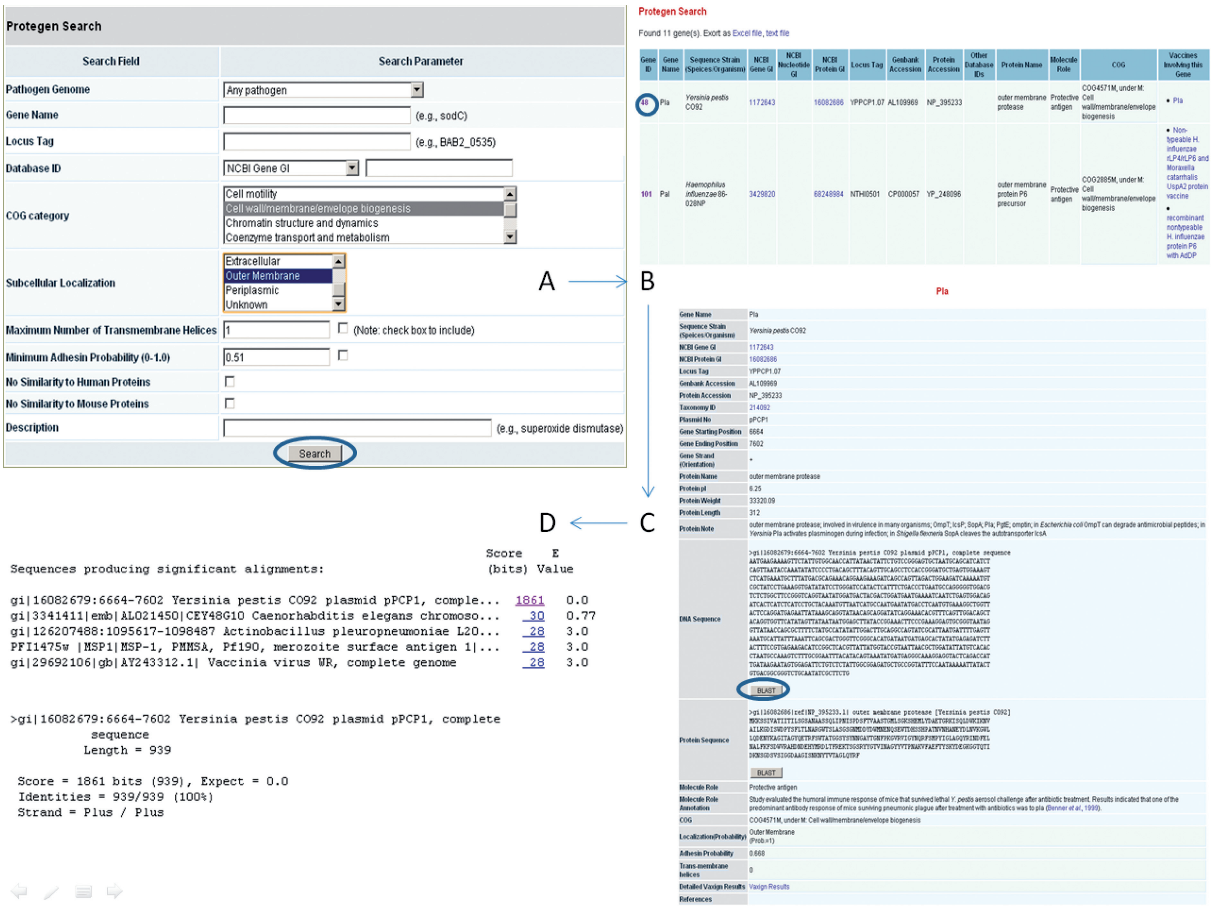
DATA TRANSFER AND DOWNLOAD

To facilitate data exchange and transfer, the information of protective antigens in Protegen is now stored in the Vaccine Ontology (VO; http://www.violinet.org/vaccineontology) (15). VO is a community-based ontology in the vaccine domain. It is developed based on the OWL format, which can be processed by many existing software programs and used for Semantic Web applications. The storage of protective antigens in VO allows development of new software programs to integrate the protective antigen data with other biomedical data and support further bioinformatics analyses. Users can also export selected protective antigen information into a Microsoft Excel document. The FASTA sequence format has frequently been used for flexible bioinformatics analyses. Protegen provides free downloads of FASTA files of merged or grouped DNA or protein sequences of protective antigens.

DISCUSSION

A few challenges have been identified throughout the curation process. One challenge is that a protective antigen is often hard to identify for many less studied pathogens (e.g. *Bordetella avium*) and less fatal diseases. In many cases, after a protective antigen was identified, the pathogen strains available from NCBI frequently did not match up with the strains used in the experimental challenge study. Unless the two strains had conserved regions of DNA for the protective antigen, we could not use these studies for inclusion in Protegen. Furthermore, many challenge studies would design a successful potential vaccine that expressed two or more distinct proteins. In our curation, these challenge studies are not typically used since it is difficult to conclude whether each individual protein induces protection *in vivo*. In case such challenge studies were included in our curation, specific notes were added to the section of experimental evidence (i.e. Molecular Annotations of antigens).

Computational and informatics technologies have been used to support immunology and vaccine development for decades (16). Prediction and analysis of immune epitopes have been emphasized from the beginning. As a result, many high quality immune epitope databases and programs, such as IEDB (17) and MHCPred (18), have been developed. Many comprehensive reviews on this topic can be found in the literature (16,19). Recently there has been a trend to study other aspects including immunogenic antigens and protective antigens. AntigenDB is a newly reported database focused on collection of immunogenic antigens (20). AntigenDB collects antigen information from selected pathogen species.



Protegen Search

Search Field	Search Parameter
Pathogen Genome	Any pathogen
Gene Name	(e.g., sodC)
Locus Tag	(e.g., BAB2_0535)
Database ID	NCBI Gene GI
COG category	Cell motility Cell wall/membrane/envelope biogenesis Chromatin structure and dynamics Coenzyme transport and metabolism
Subcellular Localization	Extracellular Outer Membrane Periplasmic Unknown
Maximum Number of Transmembrane Helices	1 □ (Note: check box to include)
Minimum Adhesin Probability (0-1.0)	0.51 □
No Similarity to Human Proteins	□
No Similarity to Mouse Proteins	□
Description	(e.g., superoxide dismutase)

Protegen Search

Found 11 gene(s). Exort as Excel file, text file

Gene ID	Gene Name	Sequence Strain (Speices/Organism)	NCBI Gene GI	NCBI Nucleotide GI	NCBI Protein GI	Locus Tag	Genbank Accession	Protein Accession	Other Database IDs	Protein Name	Molecule Role	COG	Vaccines Involving this Gene
48	Pla	*Yersinia pestis* CO92	1172643		16082686	YPPCP1.07	AL109969	NP_395233		outer membrane protease	Protective antigen	COG4571M, under M: Cell wall/membrane/envelope biogenesis	• Pla
101	Pal	*Haemophilus influenzae* 86-028NP	3429820		68248984	NTHI0501	CP000057	YP_248096		outer membrane protein P6 precursor	Protective antigen	COG2885M, under M: Cell wall/membrane/envelope biogenesis	• Non-typeable H. influenzae rLP4/rLP6 and Moraxella catarrhalis UspA2 protein vaccine • recombinant nontypeable H. influenzae protein P6 with AdDP

Pla

Gene Name	Pla
Sequence Strain (Speices/Organism)	*Yersinia pestis* CO92
NCBI Gene GI	1172643
NCBI Protein GI	16082686
Locus Tag	YPPCP1.07
Genbank Accession	AL109969
Protein Accession	NP_395233
Taxonomy ID	214092
Plasmid No	pPCP1
Gene Starting Position	6664
Gene Ending Position	7602
Gene Strand (Orientation)	+
Protein Name	outer membrane protease
Protein pI	6.25
Protein Weight	33320.09
Protein Length	312
Protein Note	outer membrane protease; involved in virulence in many organisms; OmpT; IcsP; SopA; Pla; PgtE; omptin; in *Escherichia coli* OmpT can degrade antimicrobial peptides; in *Yersinia* Pla activates plasminogen during infection; in *Shigella flexneria* SopA cleaves the autotransporter IcsA
DNA Sequence	>gi|16082679:6664-7602 Yersinia pestis CO92 plasmid pPCP1, complete sequence AATGAAGAAAAGTTCTATTGTGGCAACCATTATAACTATTCTGTCCGGGAGTGCTAATGCAGCATCATCT CAGTTAATACCAAATATATCCCCTGACAGCTTTACAGTTGCAGCCTCCACCGGGATGCTGAGTGGAAAGT CTCATGAAATGCTTTATGACGCAGAAACAGGAAGAAAGATCAGCCAGTTAGACTGGAAGATCAAAAATGT CGCTATCCTGAAAGGTGATATATCCTGGGATCCATACTCATTTCTGACCCTGAATGCCAGGGGGTGGACG TCTCTGGCTTCCGGGTCAGGTAATATGGATGACTACGACTGGATGAATGAAAATCAATCTGAGTGGACAG ATCACTCATCTCATCCTGCTACAAATGTTAATCATGCCAATGAATATGACCTCAATGTGAAAGGCTGGTT ACTCCAGGATGAGAATTATAAAGCAGGTATAACAGCAGGATATCAGGAAACACGTTTCAGTTGGACAGCT ACAGGTGGTTCATATAGTTATAATAATGGAGCTTATACCGGAAACTTCCCGAAAGGAGTGCGGGTAATAG GTTATAACCAGCGCTTTTCTATGCCATATATTGGACTTGCAGGCCAGTATCGCATTAATGATTTTGAGTT AAATGCATTATTTAAATTCAGCGACTGGGTTCGGGCACATGATAATGATGAGCACTATATGAGAGATCTT ACTTTCCGTGAGAAGACATCCGGCTCACGTTATTATGGTACCGTAATTAACGCTGGATATTATGTCACAC CTAATGCCAAAGTCTTTGCGGAATTTACATACAGTAAATATGATGAGGGCAAAGGAGGTACTCAGACCAT TGATAAGAATAGTGGAGATTCTGTCTCTATTGGCGGAGATGCTGCCGGTATTTCCAATAAAAATTATACT GTGACGGCGGGTCTGCAATATCGCTTCTG BLAST
Protein Sequence	>gi|16082686|ref|NP_395233.1| outer membrane protease [Yersinia pestis CO92] MKKSSIVATIITILSGSANAASSQLIPNISPDSFTVAASTGMLSGKSHEMLYDAETGRKISQLDWKIKNV AILKGDISWDPYSFLTLNARGWTSLASGSGNMDDYDWMNENQSEWTDHSSHPATNVNHANEYDLNVKGWL LQDENYKAGITAGYQETRFSWTATGGSYSYNNGAYTGNFPKGVRVIGYNQRFSMPYIGLAGQYRINDFEL NALFKFSDWVRAHDNDEHYMRDLTFREKTSGSRYYGTVINAGYYVTPNAKVFAEFTYSKYDEGKGGTQTI DKNSGDSVSIGGDAAGISNKNYTVTAGLQYRF BLAST
Molecule Role	Protective antigen
Molecule Role Annotation	Study evaluated the humoral immune response of mice that survived lethal *Y. pestis* aerosol challenge after antibiotic treatment. Results indicated that one of the predominant antibody response of mice surviving pneumonic plague after treatment with antibiotics was to pla (Benner *et al.*, 1999).
COG	COG4571M, under M: Cell wall/membrane/envelope biogenesis
Localization(Probability)	Outer Membrane (Prob.=1)
Adhesin Probability	0.668
Trans-membrane helices	0
Detailed Vaxign Results	Vaxign Results
References	

```
                                                                  Score    E
Sequences producing significant alignments:                      (bits) Value

gi|16082679:6664-7602 Yersinia pestis CO92 plasmid pPCP1, comple...  1861   0.0
gi|3341411|emb|AL021450|CEY48G10 Caenorhabditis elegans chromoso...    30   0.77
gi|126207488:1095617-1098487 Actinobacillus pleuropneumoniae L20...    28   3.0
PFI1475w |MSP1|MSP-1, PMMSA, Pf190, merozoite surface antigen 1|...    28   3.0
gi|29692106|gb|AY243312.1| Vaccinia virus WR, complete genome          28   3.0


>gi|16082679:6664-7602 Yersinia pestis CO92 plasmid pPCP1, complete
          sequence
         Length = 939

 Score = 1861 bits (939), Expect = 0.0
 Identities = 939/939 (100%)
 Strand = Plus / Plus
```

Figure 2. Example of protective antigen query and BLAST sequence similarity analysis. A COG category search of 'Cell wall/membrane/envelope biogenesis' in conjunction with a subcellular localization search of 'Outer Membrane' (**A**) identified 11 genes from the Protegen database, including Pla from *Yersinia pestis* strain CO92, and Pal from *Haemophilus influenza* strain 86-028NP (**B**). Clicking the Protegen antigen ID associated with Pla provided curated data including the sequence strain, NCBI Gene GI, NCBI Protein GI, protein name, NCBI taxonomy ID, DNA and protein sequences as well as other information (**C**). A BLAST sequence similarity analysis of the DNA sequence produced multiple hits with significant alignments (**D**).

One major difference between Protegen and AntigenDB is that Protegen only includes those antigens that induce protection against virulent challenge *in vivo* or stimulate a specific immune response(s) that correlate with protection. AntigenDB includes antigens that stimulate immune responses but do not necessarily induce protection or protective immune response. For example, AntigenDB includes 73 antigens from *Mycobacterium tuberculosis*. Based on our manual curation and analysis, only 14 out of the 73 antigens were experimentally verified to induce protection *in vivo* and can be classified as protective antigens in Protegen. Not all immunogenic antigens can be used for vaccine development. For example, *M. tuberculosis* LppX (21) and Hsp65 (22) are both highly immunogenic but not protective against tuberculosis. To the best of our knowledge, Protegen is the first web-based, publically available database and analysis system that targets for the curation and analysis of protective antigens.

In addition, Protegen has several other unique features compared to AntigenDB. First, Protegen manually curates and stores experimental evidence for each protective antigen. Second, Protegen covers a broad range of diseases, including infectious diseases caused by more than 100 pathogen species and many non-infectious diseases (e.g. cancers and allergies). Third, Protegen includes bioinformatics analysis of weight, pI, transmembrane domain, adhesin probability for each protective antigen stored in Protegen. Protegen also provides links to the new Vaxign vaccine design program.

VIOLIN is the first web-based vaccine database and analysis system that targets for vaccine research (6). Currently VIOLIN has included 745 vaccines and vaccine candidates for infectious diseases caused by over 100 pathogens and many non-infectious diseases. Several new programs (e.g. Vaxign, VO and Protegen) have been developed in VIOLIN. Protegen is not just one part of further VIOLIN development. Protegen is a new and relatively independent program, and the usage of protective antigens collected in Protegen may go beyond vaccine research and be used for diagnosis and immune mechanism studies.

The Protegen database can be used for many applications. For example, the protective antigens can be used as positive controls for rational vaccine design. It has long been suggested that those pathogens (e.g. *E. coli*) against which a strong antibody (B cell) response is critical, surface-exposed outer membrane proteins and secreted proteins are ideal targets for vaccine development. However, to develop vaccines against those pathogens (e.g. *Brucella* spp.) where T cell response is critical, subcellular localization may not be an issue since a T cell response could be directed to any protein target. Such a phenomenon can be verified by examining predicted subcellular localizations for all bacterial protective antigens in Protegen (Supplementary Figure S2). The Protegen protective antigens can also be used for analysis of protective immune epitopes. Protein structural analysis can significantly aid the rational design of future vaccines. One key towards a next generation rational vaccine design is the ability to apply structure-oriented bioinformatics to predict immune peptide epitopes ideal for vaccine development (12). We hypothesize that many protective antigens preserve some unique epitope information which would support further vaccine development and immune response studies. Such a hypothesis is currently under investigation. The atomic resolution of the structures of protective antigens may lead to the rational design of target epitopes as vaccine candidates.

More protective antigens will be added to Protegen. More diseases will also be covered. Currently Protegen is focused on curation of protective protein antigens. We are also in the process of annotating other types of protective antigens. For example, bacterial LPSs from many bacteria (e.g. *Brucella*) have been found to be protective antigens. Protegen is targeted to become a central and vital source of protective antigens and will support researchers in the areas of vaccinology, microbiology, and immunology with curated data and bioinformatics tools. We believe that Protegen is a timely repository and will have significant impact for vaccine research and development and other applications.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: National Institutes of Health-National Institute of Allergy and Infectious Diseases grant (R01AI081062).

Conflict of interest statement. None declared.

REFERENCES

1. Becker,K., Hu,Y. and Biller-Andorno,N. (2006) Infectious diseases: a global challenge. *Int. J. Med. Microbiol.*, **296**, 179–185.
2. Gregersen,J.P. (2001) DNA vaccines. *Naturwissenschaften*, **88**, 504–513.
3. Bousquet,J., Lockey,R. and Malling,H.J. (1998) Allergen immunotherapy: therapeutic vaccines for allergic diseases. A WHO position paper. *J. Allergy Clin. Immunol.*, **102**, 558–562.
4. Tabi,Z. and Man,S. (2006) Challenges for cancer vaccine development. *Adv. Drug Deliv. Rev.*, **58**, 902–915.
5. Tang,S. and Hewlett,I. (2010) Nanoparticle-based immunoassays for sensitive and early detection of HIV-1 capsid (p24) antigen. *J. Infect. Dis.*, **201(Suppl. 1)**, S59–64.
6. Xiang,Z., Todd,T., Ku,K.P., Kovacic,B.L., Larson,C.B., Chen,F., Hodges,A.P., Tian,Y., Olenzek,E.A., Zhao,B. *et al.* (2008) VIOLIN: vaccine investigation and online information network. *Nucleic Acids Res.*, **36**, D923–D928.
7. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
8. Gardy,J.L., Laird,M.R., Chen,F., Rey,S., Walsh,C.J., Ester,M. and Brinkman,F.S. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617–623.
9. Kall,L., Krogh,A. and Sonnhammer,E.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction: the Phobius web server. *Nucleic Acids Res.*, **35**, W429–W432.
10. Sachdeva,G., Kumar,K., Jain,P. and Ramachandran,S. (2005) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics*, **21**, 483–491.
11. He,Y., Xiang,Z. and Mobley,H.L. (2010) Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *J. Biomed. Biotechnol.*, **2010**, Article ID 297505.
12. Serruto,D. and Rappuoli,R. (2006) Post-genomic vaccine development. *FEBS Lett.*, **580**, 2985–2992.
13. Xiang,Z., Zheng,W. and He,Y. (2006) BBP: *Brucella* genome annotation with literature mining and curation. *BMC Bioinformatics*, **7**, 347.
14. Xiang,Z., Tian,Y. and He,Y. (2007) PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biol.*, **8**, R150.
15. He,Y., Cowell,L., Diehl,A.D., Mobley,H.L., Peters,B., Ruttenberg,A., Scheuermann,R.H., Brinkman,R.R., Courtot,M., Mungall,C. *et al.* (2009) *The 1st International Conference on Biomedical Ontology (ICBO 2009). Nature Precedings*, Buffalo, NY, USA. http://precedings.nature.com/documents/3553/version/1.
16. Flower,D.R. (2008) *Bioinformatics for Vaccinology*, 1st edn. Wiley-Blackwell, Oxford.
17. Vita,R., Zarebski,L., Greenbaum,J.A., Emami,H., Hoof,I., Salimi,N., Damle,R., Sette,A. and Peters,B. (2010) The immune epitope database 2.0. *Nucleic Acids Res.*, **38**, D854–D862.
18. Guan,P., Doytchinova,I.A., Zygouri,C. and Flower,D.R. (2003) MHCPred: bringing a quantitative dimension to the online prediction of MHC binding. *Appl. Bioinformatics*, **2**, 63–66.
19. De Groot,A.S. (2006) Immunomics: discovering new targets for vaccines and therapeutics. *Drug Discov. Today*, **11**, 203–209.
20. Ansari,H.R., Flower,D.R. and Raghava,G.P. (2010) AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res.*, **38**, D847–D853.
21. Lefevre,P., Denis,O., De Wit,L., Tanghe,A., Vandenbussche,P., Content,J. and Huygen,K. (2000) Cloning of the gene encoding a 22-kilodalton cell surface antigen of *Mycobacterium bovis* BCG and analysis of its potential for DNA vaccination against tuberculosis. *Infect. Immun.*, **68**, 1040–1047.
22. Pelizon,A.C., Martins,D.R., Zorzella-Pezavento,S.F., Seger,J., Justulin,L.A. Jr, da Fonseca,D.M., Santos,R.R. Jr, Masson,A.P., Silva,C.L. and Sartori,A. (2010) Neonatal BCG immunization followed by DNAhsp65 boosters: highly immunogenic but not protective against tuberculosis - a paradoxical effect of the vector? *Scand. J. Immunol.*, **71**, 63–69.