Published online 9 October 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D847–D853***
doi:10.1093/nar/gkp830

AntigenDB: an immunoinformatics database of pathogen antigens

Hifzur Rahman Ansari[1], Darren R. Flower[2] and G. P. S. Raghava[1,*]

[1]Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, India and [2]The Jenner Institute, University of Oxford, High Street, Compton, Berkshire, RG20 7NN, UK

Received July 28, 2009; Revised September 16, 2009; Accepted September 19, 2009

ABSTRACT

The continuing threat of infectious disease and future pandemics, coupled to the continuous increase of drug-resistant pathogens, makes the discovery of new and better vaccines imperative. For effective vaccine development, antigen discovery and validation is a prerequisite. The compilation of information concerning pathogens, virulence factors and antigenic epitopes has resulted in many useful databases. However, most such immunological databases focus almost exclusively on antigens where epitopes are known and ignore those for which epitope information was unavailable. We have compiled more than 500 antigens into the AntigenDB database, making use of the literature and other immunological resources. These antigens come from 44 important pathogenic species. In AntigenDB, a database entry contains information regarding the sequence, structure, origin, etc. of an antigen with additional information such as B and T-cell epitopes, MHC binding, function, gene-expression and post translational modifications, where available. AntigenDB also provides links to major internal and external databases. We shall update AntigenDB on a rolling basis, regularly adding antigens from other organisms and extra data analysis tools. AntigenDB is available freely at http://www.imtech.res.in/raghava/antigendb and its mirror site http://www.bic.uams.edu/raghava/antigendb.

INTRODUCTION

The term vaccine can be applied to all agents, either of a molecular or supramolecular nature, used to stimulate specific, protective immunity against pathogenic microbes and the disease they cause. It is clear that vaccines form the most powerful and cost-effective prophylactic therapy for infectious disease. Vaccines work to militate against the effects of subsequent infection as well as blocking the ability of a pathogen to kill its host. The availability of entire genomes corresponding to many important pathogens has instigated a new research initiative able to discover a wide array of antigens, which can act as potential vaccine candidates. Bioinformatics, in the form of comprehensive immunological databases and analysis tools, has hastened both the identification and the validation of candidate vaccines.

Previously, the principal means of antigen discovery was empirical. A live, virulent pathogen, for example, is now considered a poor candidate vaccine since despite its potent immunogenicity it is more liable to induce disease than to prevent or treat it. Thus, vaccines have, until recently, been primarily attenuated or chemically inactivated whole pathogen vaccines such as Sabin's polio vaccine or BCG. More recently, safety concerns have fostered alternate strategies for vaccine development. The most successful has focused on the antigen, acellular or subunit vaccine, which includes recombinant vaccines against hepatitis B, human papillomavirus and Haemophilus influenzae B. Subunit vaccines are typically composed of immunogenic protein or carbohydrate, such as cell wall components, or a bio-conjugated combination of both. Many antigens are highly immunogenic, while others stimulate measurable yet often weak responses, requiring boosting to perpetuate long-term protection and the addition of adjuvants (1–3).

Antigen-based subunit vaccines long ago became a prime focus on vaccine discovery. Approaches to the identification of antigens include the identification of immunodominant epitopes, assaying for enhanced correlates of protection such as antibody levels and cytokine production and testing for enhanced survival in disease challenge models. Long, cumbersome and inconclusive procedures for antigen validation currently in use, compounded by the failure to identify protective B cell or T cell mediated epitopes, are often unsuccessful, failing to detect efficiently an antigen as an efficacious vaccine candidate. Thus the identification of antigens as putative whole protein subunit vaccines is also now a key goal of immunoinformatics and computational vaccinology.

*To whom correspondence should be addressed. Tel: +91 172 2690557; Fax: +91 172 2690585; Email: raghava@imtech.res.in

The *in silico* identification of antigens offers the hope of eliciting significant responses from both humoral and cellular immune systems, far exceeding the efficacy of peptide vaccines, while avoiding the potential toxicity problems associated with whole microbe vaccines. A necessary step towards systematizing antigen discovery is the rigorous compilation and annotation of antigens.

Recently, massive factory-scale experimentation coupled to literature mining has allowed numerous functional immunology databases to emerge. Databases such as The Immune Epitope Database and Analysis Resource (IEDB), MHCBN, AntiJen and BCIPEP (4–7) are large, robust data repositories and provide copious information. Concerning antigens, however, these databases, although replete with information concerning individual B cell epitopes, T cell epitopes and Major Histocompatibility Complex (MHC) binding peptides, remain otherwise partial and incomplete. Their focus is on the epitope, and not the antigen. There are many antigens, for which specific epitope or MHC binding information is not currently available, yet many such antigens are known experimentally to induce either or both innate or adaptive immune responses. Such antigens—or similar pathogenic proteins—might prove useful in vaccine design. These antigens require urgent and rigorous cataloguing.

In order to address this pressing issue, the present work describes the database AntigenDB. It is a specialized, value-added database of antigens derived from pathogenic organisms. This resource is intended to be a repository for all experimentally determined antigens, irrespective of whether such an antigen is associated within the extant knowledge-base with known epitope data. The database is freely accessible through a web browser at http://www.imtech.res.in/raghava/antigendb/.

SYSTEM AND METHODS

Database construction and architecture

Experimentally validated antigens were collected from the literature (PubMed: http://www.ncbi.nlm.nih.gov/pubmed/; ScienceDirect: http://www.sciencedirect.com/). Additional information about these antigens was collected from various public databases including IEDB, MHCBN, AntiJen and BCIPEP. We developed PERL scripts to extract sequence, structural, functional and gene expression information from SwissProt, GenBank, PDB (Protein Data Bank) and GEO databases (8–11). AntigenDB is built on a SUN systems T-1000 under Solaris 10.0 environment. The front-end was developed using HTML and the backend was developed using PostgreSQL, a relational database management system. All common gateway interface (CGI) and database interfacing scripts were written in the PERL programming language. The architecture of AntigenDB database is shown in Figure 1a and b.

Organization of data

AntigenDB collects and compiles comprehensive information concerning antigens. Most of the antigens in AntigenDB come from the genus Mycobacterium, Plasmodium and the Influenza A virus (Figure 2). The up-to-date status of the database is available at url: http://www.imtech.res.in/raghava/antigendb/info.html.
Data for each antigen can be categorized as primary data (antigen sequence and structural information), B cell epitope (epitope sequence and antibody information), T cell epitope (T helper and T killer cell epitopes, MHC I/II binding and TAP binding and cleavage sites information), function (cellular location, function, functional sites and similarity with host and pathogens), Gene expression (nucleotide sequence, codon frequency and expression profiles) and different types of Post-translational modification (PTM) associated with antigens. Each antigen is assigned a unique entry number and information is divided into six tables; each table providing unique information.

General information. The main or primary table contains general information about antigens. This table has the following main fields: (i) antigen name: the name of the antigen with its synonyms; (ii) antigen type: whether it is protein, carbohydrate or lipid; (iii) amino acid sequence of protein antigen and (iv) source organism of origin, with a link to the NCBI taxonomy database.

Structural information. Within the database, detailed crystal structure information is available for 290 molecules out of 504 antigens. This information is supplemented by the OCA web browser (http://oca.weizmann.ac.il), with surface accessibility provided by ASAView tools (12). We also link to the Swiss-model repository database (13), which provides hypothetical structures for unsolved protein using protein-modeling techniques. We also provide secondary structure information in the form of percent content as calculated by DSSP (14).

B-cell epitope. A principal challenge for immunology is to identify antigenic regions, responsible for simulating B-cells, also called B-cell epitopes (15). We have collected B-cell epitopes reported for antigens available in AntigenDB (Figure S1). This table has the following major fields: (i) the capability of antigen to induce B cell or humoral immune responses; (ii) specific B-cell epitopes within the antigen; (iii) the antibodies that recognize these epitopes and (iv) PTMs associated with such epitopes. Most of the data for this table has been obtained from the primary literature or from a secondary source (BCIPEP and IEDB). Many antigens have no known B-cell epitopes; these antigens are not covered by B-cell epitope databases.

T-cell epitope. Most extant vaccines are mediated by antibodies. However, responses to diseases without effective vaccines are largely mediated by cellular—not humoral—immunology. Thus, to develop subunit vaccines one needs to identify T-cell epitopes within an antigen. Figure S2 shows the distribution of T cell epitopes in antigenDB; most antigens have less than 10 T-cell epitopes. The T-cell epitope table contains the following main fields: (i) immunogenicity induced due to T

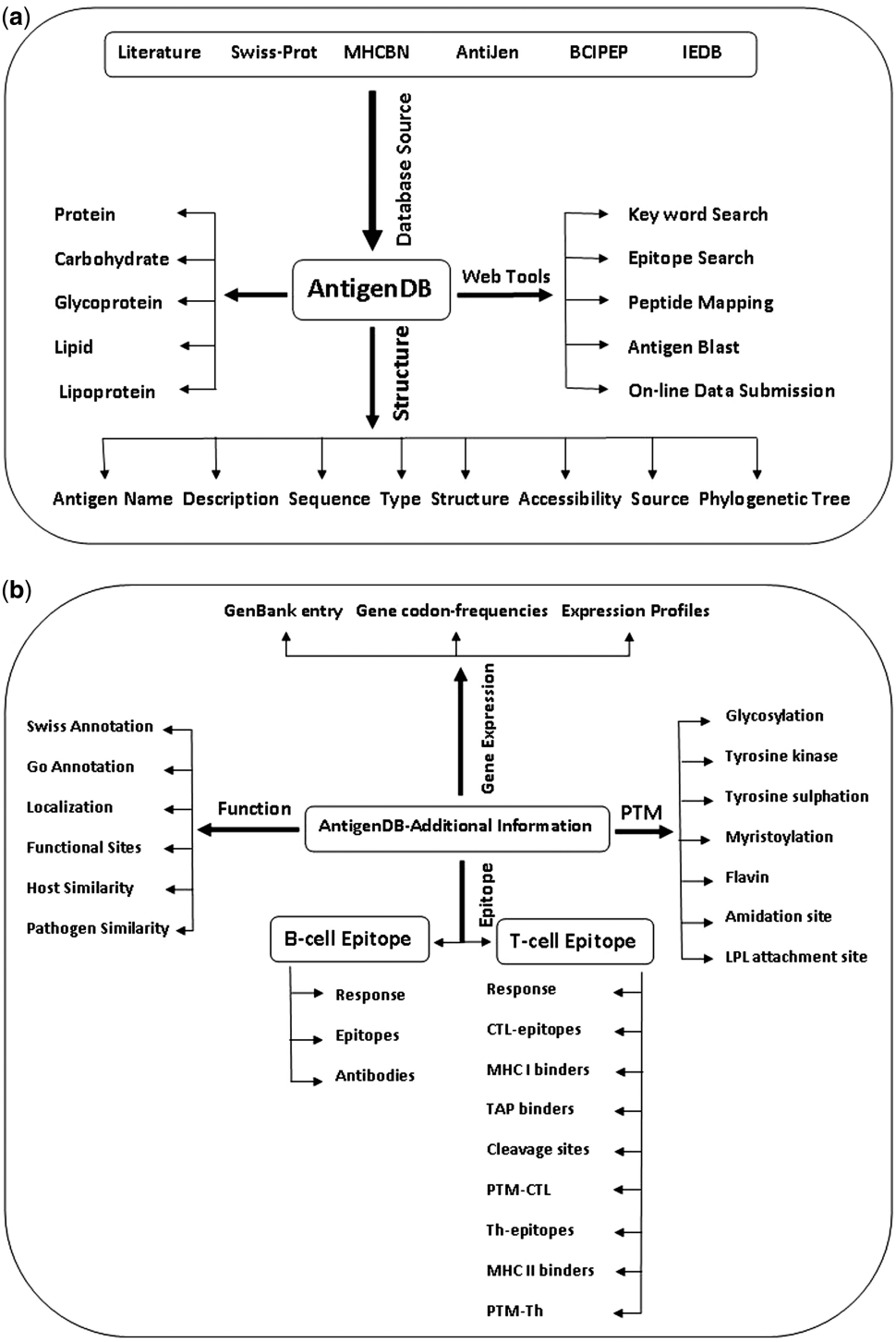

Figure 1. (**a**) Overall architecture of AntigenDB. (**b**) Detailed information for each antigen entry.

helper or Cytotoxic T cells; (ii) Helper or cytotoxic T cell epitopes; (iii) PTMs associated with these epitopes; (iv) MHC I/II binders with experimental IC50 values if available; (v) TAP binders mapping and (vi) mapping of cleavage sites. Data for this table was obtained from the literature, and from the MHCBN, IEDB and AntiJen databases. Both B and T-cell epitopes are linked directly to other epitope databases, such as the IEDB. There are ~358 antigens for which there is no epitope information; this means these antigens are not covered in existing epitopes databases. There are about 95 antigens of genus Mycobacterium, alone which are shown to induce immune response, but for which no epitope has yet been identified. This demonstrates the importance of AntigenDB, which covers experimentally-validated, protective antigens not covered in epitope-oriented databases.

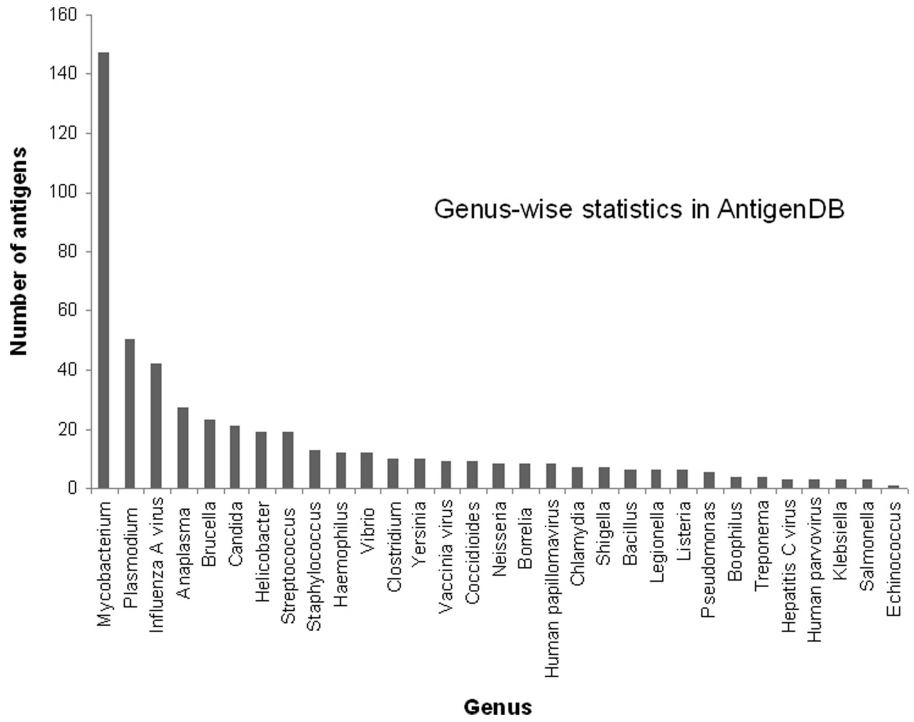

Figure 2. Antigen distribution in AntigenDB database.

Functional information. The function of an antigen determines potential candidacy as a vaccine. If the antigen is involved in house-keeping or is present at the cell surface and easily available to surveillance by the host immune system then the probability that an antigen will be suitable vaccine candidate increases. Therefore, there is an imperative need to have insight into the function and subcellular localization of each antigen. The function table provides the following information: (i) functional annotation of antigens using SwissProt and GO databases; (ii) Cellular localization (secreted, cytoplasmic and membrane bound) as described in DBSubLoc and PSORTdb (16,17) databases and (iii) functional sites and domains obtained from InterPro (18). The development of a better vaccine requires knowledge of the similarity between candidate antigen and others derived from the host or other pathogenic organisms. If an antigen is similar to one or more host proteins there exists the possibility of autoimmune responses; less similarity with other pathogens is also advantageous and would qualify the antigen as a better vaccine candidate. Therefore, we provide a BLAST (19) results for each antigen with pathogenic and human proteins.

Gene expression. Suitable antigens are often highly expressed and thus optimally available to host surveillance. Expression of any protein is largely dependent on the codon usage of that organism (20). Therefore, AntigenDB provides information about: (i) Gene sequence as obtained from GenBank; (ii) codon frequency of genes; (iii) codon bias calculated using Graphical codon usage analyzer (GCUA) (21) and (iv) expression profiles of antigen obtained from databases such as GEO (11).

Post-translational modification. PTMs affect the expression and function of antigens. Therefore different types of PTMs are covered in the database. The major PTMs covered include: (i) N/O/C/S-Glycosylation; (ii) Phosphorylation; (iii) Amidation site; (iv) N-Myristoylation; (v) Tyrosine Sulfation and (vi) Methylation and other PTMs. These are compiled from the literature as well as specialized databases such as dbPTM, PhosphoELM, and RESID (22–24).

IMPLEMENTATION

Currently, AntigenDB contains an extensive collection of proteins, glycoproteins and lipoproteins (in excess of 500), extracted from 44 important pathogenic species. This covers following major genuses; Mycobacterium, Influenza A virus, Helicobacter, Bacillus, Brucella, Clostridium, Hepatitis, Plasmodium, Streptococcus, Yersinia and Vibrio (Figure 2).

AntigenDB entries are cross-linked to a variety of key databases, such as SwissProt, GenBank, IMGT (25), SYFPEITHI (26), AntiJen, DBSubLoc, PSORT and the PDB. In AntigenDB, different types of useful web tools have been provided. There are many tools integrated into AntigenDB for the extraction and analysis of antigens. These include searching the AntigenDB database, the analysis of antigen data through mapping, and data submission.

Data search and analysis

AntigenDB is complemented by an array of tools, which facilitate further analysis of antigens (Figure 3). AntigenDB is user-friendly and can be searched using

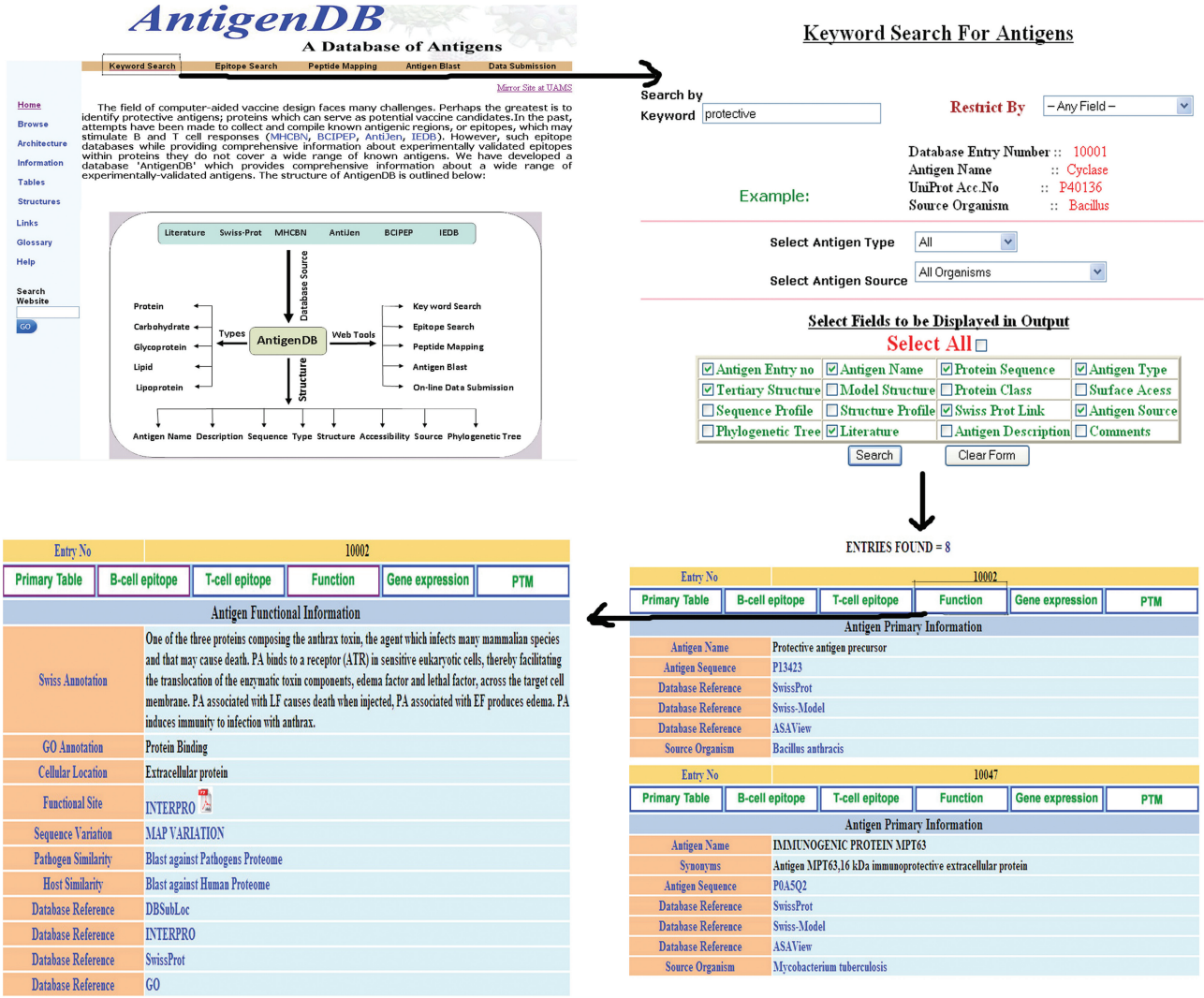

Figure 3. Flow of data searching and result reporting. From AntigenDB home page to data search using keyword 'protective' and display of results.

antigen name, SwissProt accession no, PDB ID or organisms name. Users of AntigenDB can quickly extract useful information from the database in two different ways: (i) via Keyword search and (ii) via Browsing.

Keyword search. The keyword search allows a user to search for antigens using the five digits AntigenDB entry number, antigen name, SwissProt accession number or by entering a source microorganism to list all the available antigens present in that pathogen. A query can be filtered further by selecting the antigen type among the protein, carbohydrate and lipid antigens or by selecting the specific organism from the drop down menu. A query returns a tabular output whose format depends upon the fields selected initially, with an option of exporting the search result as a data file.

Browsing. Browing is a very useful tool in database, which can be searched easily by selecting organism of interest and output fields. It returns all the available antigens present for that pathogen.

Antigenic BLAST

The Antigenic blast page provides users with an opportunity to search a query protein sequence against the AntigenDB database for the purposes of sequence comparison. The standalone WWW-BLAST program is carefully integrated into AntigenDB for this purpose, and has a customizable weight matrix and E score threshold.

Peptide mapping

The peptide-mapping tool allows users to ask whether a query protein contains any already known antigenic epitopes. The tool scans all the epitopes present in the AntigenDB database against the query protein and returns the starting and ending position of experimentally defined epitopes corresponding to the query protein

sequence. It also links to the AntigenDB antigens, from where returned epitopes were derived.

Epitope search

The epitope search enables users to ask whether a query epitope sequence is present in known antigens within the database. This tool searches for exact or similar epitopes in the database. It returns AntigenDB antigens in which such epitopes are present.

Data submission

The online data submission tool allows users to submit a newly identified antigen not present in the AntigenDB database. Submitted antigens are included in AntigenDB after validation.

DISCUSSION AND CONCLUSIONS

AntigenDB should prove of value to a variety of researchers: immunoinformaticians developing new and enhanced methods for the prediction of antigens (27,28); vaccine development scientists searching for antigens in newly defined pathogens or novel candidate vaccine antigens in well known pathogenic microorganisms and microbiologists analyzing virulence mechanisms in pathogenic microorganisms, amongst others.

AntigenDB is a repository for experimentally determined antigens. It is the first of its kind; in contrast to existing epitope-orientated database, AntigenDB emphasizes experimentally-verified antigens, irrespective of whether epitope information is known or unknown. The database provides useful analysis tools able to search and map an unknown protein for similarity to known antigens or experimentally determined epitopes. We have and continue to undertake exhaustive literature searches in order to cover as many antigens as possible, yet it remains possible that certain antigens are missing from the database. Initially, we intended to cover all pathogenic species, but realized quickly that this was an impractical approach. Instead, we have selected species with an impact on global health and antigens, which are available in the literature and other sources. Several species were also selected whose close relatives were already present in the database. To enable faster database accesss, we have created a mirror site at http://www.bic.uams.edu/raghava/antigendb/. The current database mainly contains protein antigens. This is influenced by the easy availability of such antigens within the current literature.

We anticipate that this thorough and comprehensive database will be extended to effective completeness, and then maintained and its content expanded, with constantly enhanced search and analysis features added on a rolling basis. For example, where available we will add experimental data derived from the literature and archived microarray experiments relating to the expression of antigens in the next release of AntigenDB. Certain important viral pathogens—such as Hepatitis A, Hepatitis B, Hepatitis E, Herpes simplex virus and Human adenovirus—and other antigen types will be included in future releases of the database.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

While this project has not been funded directly by it, we gratefully acknowledge the Biotechnology and Biological Sciences Research Council (BBSRC) grant India PA 1713. Dr Flower is a Jenner Investigator supported by the Jenner Foundation.

FUNDING

Council of Scientific and Industrial Research (CSIR). Funding for open access charge: CSIR.

Conflict of interest statement. None declared.

REFERENCES

1. Pashine,A., Valiante,N.M. and Ulmer,J.B. (2005) Targeting the innate immune response with improved vaccine adjuvants. *Nat. Med.*, **11**, S63–S68.
2. Lata,S. and Raghava,G.P.S. (2008) PRRDB: a comprehensive database of Pattern-Recognition Receptors and their ligands. *BMC Genomics*, **9**, 180.
3. Bayry,J., Tchilian,E.Z., Davies,M.N., Forbes,E.K., Draper,S.J., Kaveri,S.V., Hill,A.V., Kazatchkine,M.D., Beverley,P.C., Flower,D.R. *et al.* (2008) In silico identified CCR4 antagonists target regulatory T cells and exert adjuvant activity in vaccination. *Proc. Natl Acad. Sci USA*, **105**, 10221–10226.
4. Peters,B., Sidney,J., Bourne,P., Bui,H.H., Buus,S., Doh,G., Fleri,W., Kronenberg,M., Kubo,R., Lund,O. *et al.* (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol.*, **3**, e91.
5. Bhasin,M., Singh,H. and Raghava,G.P. (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*, **19**, 665–666.
6. Toseland,C.P., Clayton,D.J., McSparron,H., Hemsley,S.L., Blythe,M.J., Paine,K., Doytchinova,I.A., Guan,P., Hattotuwagama,C.K. and Flower,D.R. (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res.*, **1**, 4.
7. Saha,S., Bhasin,M. and Raghava,G.P. (2005) Bcipep: a database of B-cell epitopes. *BMC Genomics*, **6**, 79.
8. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
10. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F. Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
11. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.

12. Ahmad,S., Gromiha,M., Fawareh,H. and Sarai,A. (2004) ASAView: database and tool for solvent accessibility representation in proteins. *BMC Bioinformatics*, **5**, 51.
13. Kopp,J. and Schwede,T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.*, **32**, D230–D234.
14. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
15. Saha,S. and Raghava,G.P.S. (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins*, **65**, 40–48.
16. Guo,T., Hua,S., Ji,X. and Sun,Z. (2004) DBSubLoc: database of protein subcellular localization. *Nucleic Acids Res.*, **32**, D122–D124.
17. Rey,S., Acab,M., Gardy,J.L., Laird,M.R., deFays,K., Lambert,C. and Brinkman,F.S. (2005) PSORTdb: a protein subcellular localization database for bacteria. *Nucleic Acids Res.*, **33**, D164–D168.
18. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
19. 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.
20. Manoj,S., Babiuk,L.A. and van Drunen Littel-van den Hurk,S. (2004) Approaches to enhance the efficacy of DNA vaccines. *Crit. Rev. Clin. Lab. Sci.*, **41**, 1–39.
21. McInerney,J.O. (1998) GCUA: general codon usage analysis. *Bioinformatics*, **14**, 372–373.
22. Lee,T.Y., Huang,H.D., Hung,J.H., Huang,H.Y., Yang,Y.S. and Wang,T.H. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.*, **34**, D622–D627.
23. Diella,F., Cameron,S., Gemund,C., Linding,R., Via,A., Kuster,B., Sicheritz-Ponten,T., Blom,N. and Gibson,T.J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.
24. Garavelli,J.S. (2004) The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics*, **4**, 1527–1533.
25. Lefranc,M.P., Giudicelli,V., Ginestoux,C., Jabado-Michaloud,J., Folch,G., Bellahcene,F., Wu,Y., Gemrot,E., Brochet,X., Lane,J. *et al.* (2009) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.*, **37**, D1006–D1012.
26. Rammensee,H., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanovic,S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.
27. Doytchinova,I.A. and Flower,D.R. (2007) Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. *Vaccine*, **25**, 856–866.
28. Doytchinova,I.A. and Flower,D.R. (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, **8**, 4.