Published online 3 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D455–D460***
doi:10.1093/nar/gkm925

IDBD: Infectious Disease Biomarker Database

In Seok Yang[1], Chunsun Ryu[2], Ki Joon Cho[1], Jin Kwang Kim[1], Swee Hoe Ong[3], Wayne P. Mitchell[4,5], Bong Su Kim[2], Hee-Bok Oh[2] and Kyung Hyun Kim[6,*]

[1]Department of Life Sciences & Biotechnology, School of Life Sciences & Biotechnology, Korea University, Seoul, Korea, [2]Center for Infectious Diseases, Korea National Institute of Health, Seoul, Korea, [3]Genome Institute of Singapore, A*STAR, 60 Biopolis Street, Singapore, [4]Experimental Therapeutics Center, 31 Biopolis Street, Singapore, [5]Division of Information Systems, School of Computer Engineering, Nanyang Technological University, Singapore and [6]Department of Biotechnology & Bioinformatics, College of Science & Technology, Korea University, Chungnam, Korea

Received August 14, 2007; Revised September 14, 2007; Accepted October 10, 2007

ABSTRACT

Biomarkers enable early diagnosis, guide molecularly targeted therapy and monitor the activity and therapeutic responses across a variety of diseases. Despite intensified interest and research, however, the overall rate of development of novel biomarkers has been falling. Moreover, no solution is yet available that efficiently retrieves and processes biomarker information pertaining to infectious diseases. Infectious Disease Biomarker Database (IDBD) is one of the first efforts to build an easily accessible and comprehensive literature-derived database covering known infectious disease biomarkers. IDBD is a community annotation database, utilizing collaborative Web 2.0 features, providing a convenient user interface to input and revise data online. It allows users to link infectious diseases or pathogens to protein, gene or carbohydrate biomarkers through the use of search tools. It supports various types of data searches and application tools to analyze sequence and structure features of potential and validated biomarkers. Currently, IDBD integrates 611 biomarkers for 66 infectious diseases and 70 pathogens. It is publicly accessible at http://biomarker.cdc.go.kr and http://biomarker.korea.ac.kr.

INTRODUCTION

Infectious diseases remain among the leading causes of death and disability worldwide. About 15 million (>25%) of 57 million annual deaths are estimated to be related directly to infectious diseases (1). Newly emerging and re-emerging infectious diseases constitute an urgent and ongoing threat to public health throughout the world. The discovery of acquired immune deficiency syndrome (AIDS) has led to renewed appreciation of the consequences of the emergence of infectious diseases. Severe acute respiratory syndrome (SARS) emerged in southern China in 2002 and has had a profound impact on public health (2). Influenza viruses possess evolutionary agility and the capacity to jump between fowl, farm animal and human species (3). Just as troubling are chronic infections, which create persistent social and economic havoc. Recent studies have shown that the burden of morbidity and mortality associated with certain infectious diseases falls primarily on infants and young children (4), with long-term social and economic consequences.

Surveillance and early response to infectious diseases depend on rapid clinical diagnosis and detection, which, if in place, are able to ameliorate suffering and economic loss. Biomarkers, molecules that can be sensitively measured in the human body, are by definition potentially diagnostic. The efficacy of biomarkers to infectious diseases lies in their capability to provide early detection, establish highly specific diagnosis, determine accurate prognosis, direct molecular-based therapy and monitor disease progression (5). They are increasingly important in both therapeutic and diagnostic processes, with high potential to guide preventive interventions. Vast resources have been devoted to identifying and developing biomarkers that can help determine the treatments for patients. Furthermore, there is growing consensus that multiple markers will be required for most diagnoses, while single markers may serve in only selected cases. Despite intensified interest and research, however, the rate of development of novel biomarkers has been falling (6), suggesting that a resource that leverages existing data is overdue. At present the databases containing information about biomarkers are focused predominantly on cancer: early detection research network (7), gastric cancer knowledgebase (8), integrated cancer biomarker information system (9) and database for cancer, asthma and autism for children's study (10). Even here, although

*To whom correspondence should be addressed. Tel: 82-2-3290-3444; Fax: 82-2-3290-3945; Email: khkim@korea.ac.kr
Correspondence may also be addressed to Hee-Bok Oh. Tel: 82-2-355-5601; Fax: 82-2-382-4891; Email: hboh@nih.go.kr

Table 1. Current number of biomarker entries in IDBD

Disease groups	Disease	Pathogen	Biomarker
Gastrointestinal infection	14	14	107
Respiratory infection	16	18	154
Neurological infection	2	1	14
Urogenital infection[a]	9	10	46
Viral hepatitis[a]	5	5	10
Hemorrhagic fever	4	4	37
Zoonosis	7	6	87
Arbovirus infection	5	5	110
Antibiotics resistance[a]	6	6	83
Bioterrorism	8	8	83
Total[b]	66	70	611

[a]Diseases, pathogens and biomarkers without overlapping with other groups.
[b]Total number of diseases, pathogens and biomarkers without overlapping.

15–20% of cancers are linked to infectious diseases and chronic infection causes cancer (11), no systematic effort has been described for integrating information from the cancer biomarker and the infectious disease domains.

In order to advance our understanding of biomarkers and the roles in early infection processes, we have developed an integrated user-friendly relational database that catalogs putative and validated biomarkers relates them to infectious diseases processes. In addition, we have added value by hosting various bioinformatics tools that can be used to analyze and visualize the biomarker data. This freely accessible resource will be a valuable research tool and a contribution to improved public heath.

OVERVIEW OF THE DATABASE

Infectious Disease Biomarker Database (IDBD) introduces a community annotation database of biomarkers, with interfaces for users to directly edit their content and to keep track of editing history, thus capturing community knowledge and expertise. It was designed to collect, store and display information about biomarkers, conjoined to research tools for sequence and structural analyses of the data. IDBD currently includes information on 611 biomarkers from 66 infectious diseases and 70 pathogens (Table 1). Biomarkers were classified according to detection, diagnosis, pathogen typing and virulence factor for clinical or epidemiological studies and application. Validated biomarkers were regarded as representative markers for experimental verification such as detection and diagnosis of infectious diseases in the reference and specialized laboratories or in scientific literatures. Potential biomarkers were defined as those frequently cited in the context of detection and diagnosis of infectious disease in recently published research journals. The correct assignment of biomarker subtypes and the evaluation of potential or validated biomarkers critically depend on expert group in infectious disease research fields. The IDBD data are updated and modified on a regular basis by a curation team, composed of researchers at the Center for Infectious Diseases and the Center for Immunology and Pathology in the Korea National Institute of Health (KNIH) in Seoul. The content in IDBD is open and freely accessible to the general public, and IDBD is a part of the National Disease Biomarker Bank project, an integrated framework for identifying, collecting, distributing and managing of biomarkers, which is being developed at KNIH.

DATABASE DESIGN AND CONTENTS

IDBD primarily consists of the three main tables: disease, pathogen and biomarker arranged in one Oracle schema (12). Infectious diseases are divided into 10 subgroups according to the infection site and the unique features of the pathogen or disease: gastrointestinal infection, respiratory infection, neurological infection, urogenital infection, viral hepatitis, hemorrhagic fever, zoonosis, arbovirus infection, antibiotics resistance and bioterrorism. In total, 10 disease subgroups contain 66 diseases (Table 1). Each disease at IDBD is characterized by a number of attributes such as general information, pathogen, infection, symptom, diagnosis, treatment and prevention. Pathogens are grouped into bacteria, virus, fungi and parasite, currently comprising 70 pathogens of mostly bacteria and virus. Each pathogen is characterized by general information, disease, biomarker list and related infection. Validated and potential biomarker entries are divided into three categories of detection/diagnosis, pathogen typing and virulence factor, and users can then access the data according to these criteria (Figure 1A).

The database contains approximately 8–9 biomarkers per pathogen (Table 1), comprising proteins, nucleic acids, carbohydrates and small molecules. Each biomarker contains a number of categories of information: general information, detection, mechanism, pathogen link, sequence information, NCBI link, secondary structure, tertiary structure, PDB link and reference. Biological functions or roles of biomarkers are also included, if available. Sequences are obtained from databases of protein and nucleic acid sequences of the National Center for Biotechnology Information (NCBI) (13). Information of secondary and tertiary structures is obtained from PDBsum at the European Bioinformatics Institute (EBI) (14) and Protein Data Bank (PDB) (15), respectively.

DATA RETRIEVAL

Biomarker data can be retrieved efficiently through establishment of entry portals for search functions. Users can access biomarker records from the front page by clicking Biomarker at the top menu, which then allows browsing of biomarkers in alphabetical order (Figure 1B). Two different search options are provided: Simple Search can query the complete database by selecting one or all three groups of biomarkers (protein, nucleic acid and carbohydrate). Only one small molecule, a catechol siderophore, is currently deposited. Specific database queries can be defined using the Complex Search feature (Figure 1C), where fields of interest (class of pathogen, molecular type, pathogen name, biomarker name and NCBI accession number) can be selected from

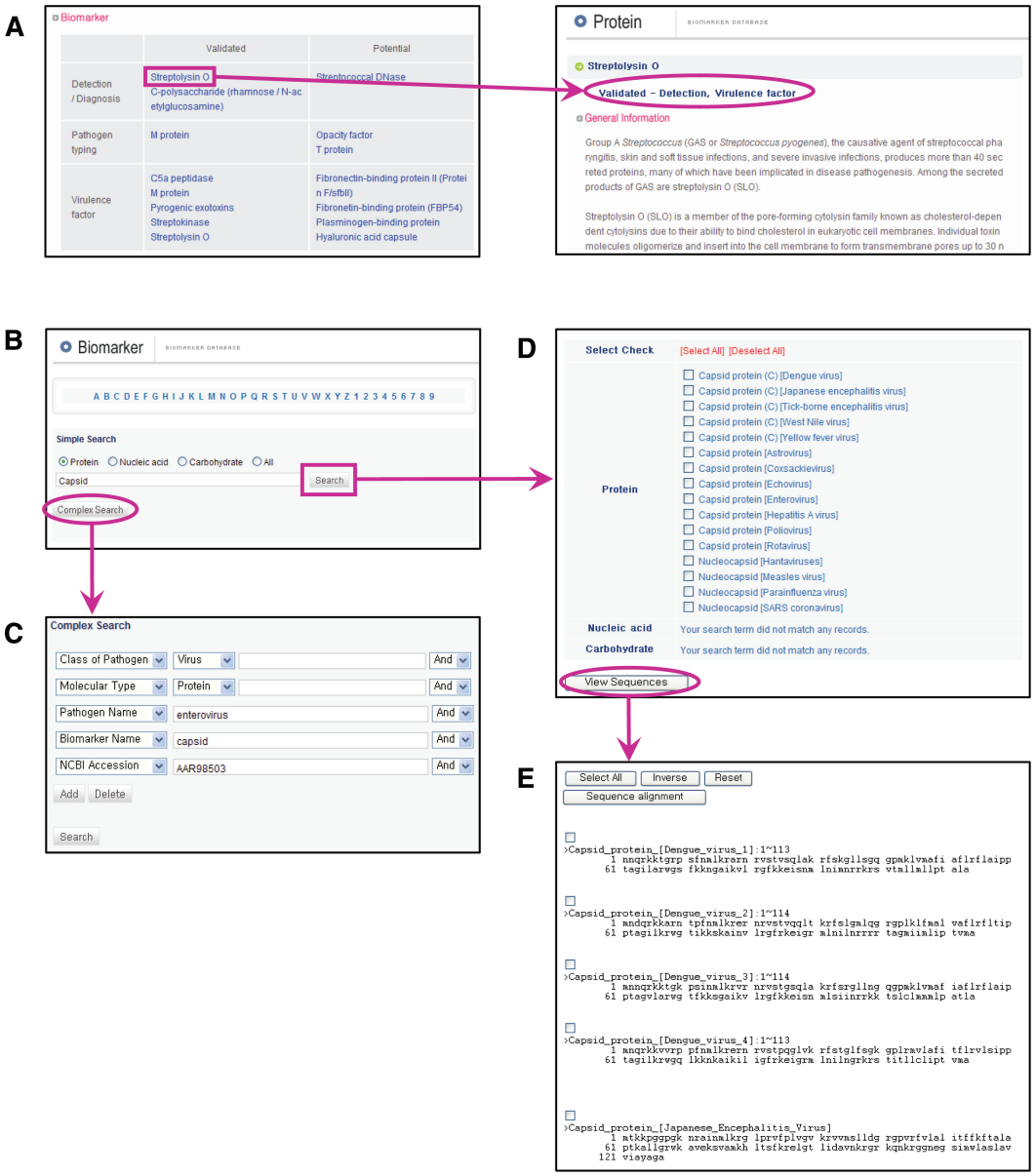

Figure 1. A screenshot of IDBD showing (**A**) the list of biomarkers at the page of pathogen, classified according to detection/diagnosis, pathogen typing and virulence factor, and the biomarker subtype page, (**B**) the list of biomarkers in alphabetical order, with two search options: Simple Search and Complex Search, (**C**) Complex Search page, (**D**) search results and (**E**) retrieved sequences viewed in a separate window.

pull-down menus. Users can iterate and append the fields in the pull-down menus by clicking the Add or Delete button and in turn combining by AND or OR operators. Both Simple and Complex Search returns a result list of distinct biomarkers that matches the search criteria (Figure 1D). The listed entries of biomarkers are linked to detailed information (general information, detection, mechanism, etc.), which can be viewed by clicking the name of each biomarker. The sequence information in the displayed list can be retrieved in a separate window (Figure 1E), as amino acid sequence in single letter abbreviation (protein) or nucleotide sequence in FASTA format (nucleic acids) of the selected biomarkers.

The complete contents of IDBD as well as other Web resources can be searched via its user-friendly interface. Three different search options are provided: Internal, External and News Search (Figure 2A). Internal Search can query via either title or content search which extracts textual content stored in the IDBD database including disease, pathogens and biomarkers (Figure 2B). In the case of External Search, users can submit a single query to the NCBI PubMed, NCBI Entrez sequence database, PDB structure database and Centers for Disease Control and Prevention (CDC) (Figure 2C). Details of published articles, sequence and structural data or disease and pathogen information can be viewed in abstract terms and sorted with regard to the information items. More detailed

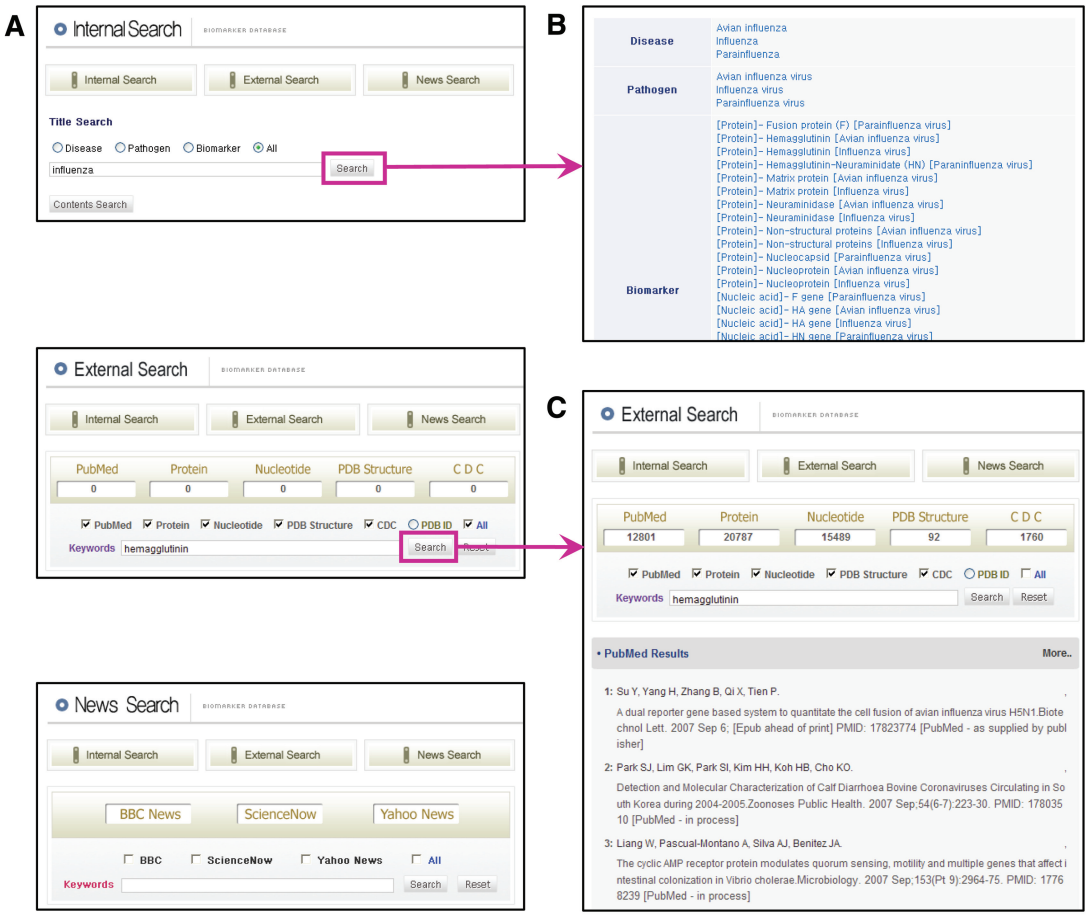

Figure 2. Search examples showing (**A**) three search options: Internal, External and News Search, (**B**) Internal Search results and (**C**) External Search results.

information can be retrieved by clicking the posting title of sorted search data. News Search provides accurate and timely information on the outbreaks and sporadic cases of any infectious diseases. Users can search current news or archives by entering keywords from scientific resources including BBC News, ScienceNow from Science magazine and Yahoo! News. More detailed information can be displayed by clicking on the title of news articles sorted by date.

DATA ANALYSIS

On the biomarker list page, IDBD allows access to sequence and structure analysis tools by clicking 'Data analysis new window' on the left-hand navigation bar. The user can select sequences of interest obtained from biomarker search (Figure 1E), prepare an input data by clicking 'Sequence alignment', and conduct multiple sequence alignment by direct submission or upload a file of the chosen sequences to CLUSTALW tool of EBI (16). On the result page (Figure 3A), the alignment can then be exported for phylogenetic tree construction. The sequence analysis also includes standard BLAST services (blastn and blastp) on external network connectivity (17), pairwise distance and synonymous—non-synonymous ratio analysis. In the case of structure analysis, users may submit an amino acid sequence or multiple sequences aligned to a set of methods for secondary (PSA) and tertiary structure (Geno3d) prediction (18,19). The predicted tertiary structures can be modeled by using Jmol either with a PDB file or PDB ID, if available (Figure 3B).

FUTURE DIRECTIONS

Pathogen-specific biomarkers can provide information necessary for diagnosis, detection and treatment of various infectious diseases. Recent concerns about bioterrorism and emerging infectious diseases have led to a new focus on the development of biomarkers, and molecular diagnostics are growing fast in infectious diseases. Currently, there are 611 biomarkers in IDBD among which 239 are validated ones. The content of biomarkers in IDBD is expanding rapidly, and our goal is to collect a complete dataset of validated or potential biomarkers used in the detection of infectious diseases and to generate a knowledgebase that would be a valuable tool for users interested in the discovery of infectious disease biomarkers. The main challenge in the future is to keep IDBD up to date with the growing number of biomarkers experimentally verified and published in peer

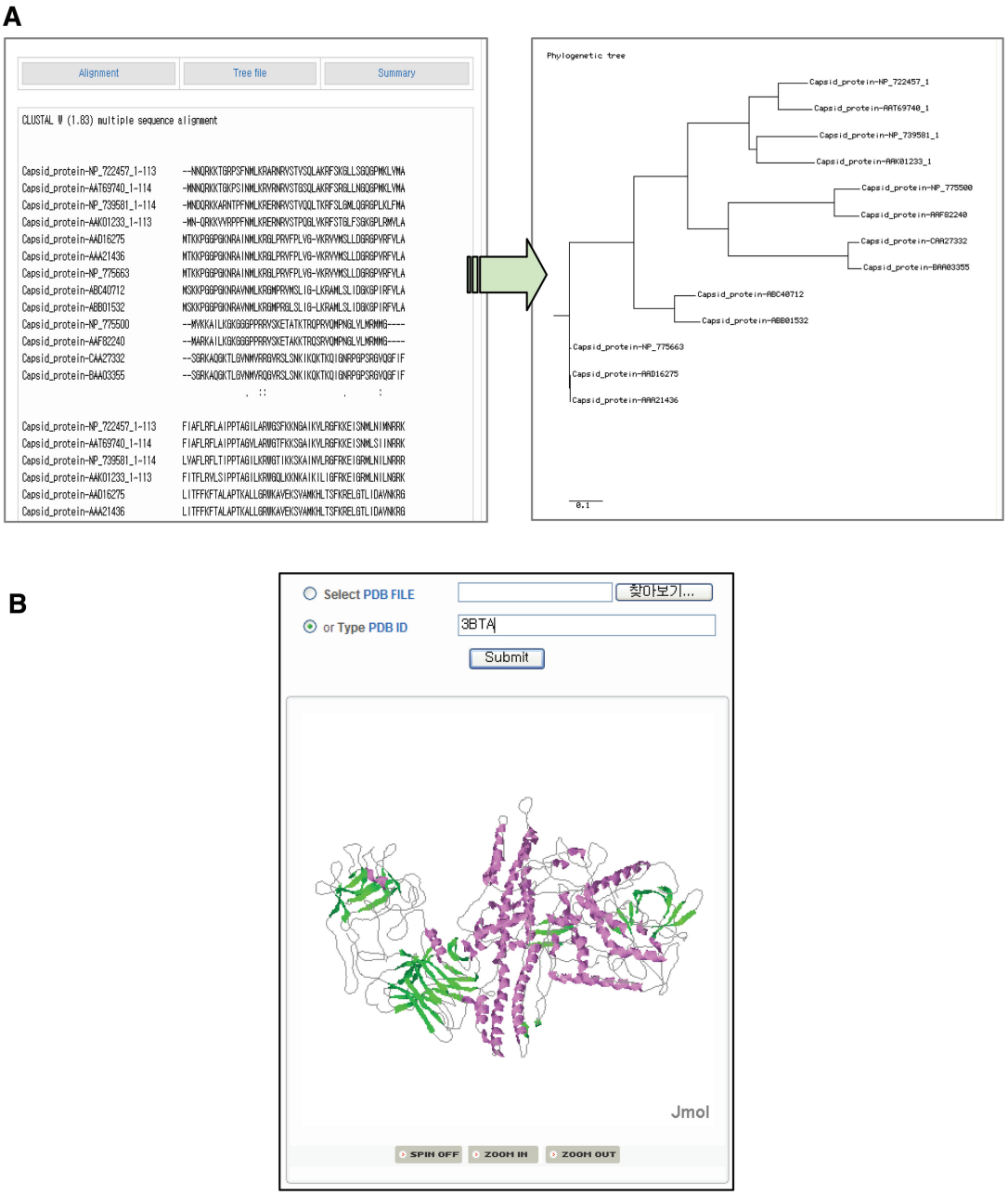

Figure 3. The sequence and structure analysis tools showing (**A**) analysis results using selected sequences of biomarker data: sequence alignment and phylogenetic tree, and (**B**) predicted 3D structure of a biomarker viewed with Jmol.

reviewed journals. We will thus implement text-mining support for database curation in the near future. Toward this goal, a network of biomarker expert groups at the Center for Infectious Diseases and the Center for Immunology and Pathology at KNIH and advisory committee outside KNIH has been formed to coordinate database management and validation of novel biomarkers.

Another challenge is to provide IDBD with regional epidemiological trends of some infectious diseases. Epidemiological surveillance is key to control and monitor diseases such as cholera, shigellosis, typhoid fever, paratyphoid fever, rabies, etc. Since information on infectious agents is essential in preventing and controlling the spread of infectious diseases, it is necessary to collect, analyze and publish the relevant information on a regional scale, to the benefit of the researchers studying infectious diseases as well as the public. New insights into innate immunity initiated by host–pathogen interaction are changing the way we think about pathogenesis of infectious diseases. Different approaches are employed for the characterization of immune responses by evaluating the epitopes recognized by antigen-specific receptors of immune systems. We intend to construct additional tools useful for molecular immunological and etiological applications. IDBD will collect intrinsic epitope features, seasonal patterns of sequences and epitope responses to

T and B cells, and develop software for epitope analysis and prediction.

USER MANAGEMENT

The IDBD management system allows users to access the biomarker database without registration. However, user registration is required for adding and editing database contents, and user support can be obtained by e-mailing graduate@korea.ac.kr. Readers are encouraged to contact us if they wish to provide new data for inclusion in IDBD, assist with curation or have any suggestions for improvements.

IMPLEMENTATION

IDBD was developed as a relational database using Oracle 10g applications (12) on the Windows operating system. Two open source softwares, the Apache HTTP Server and Apache Tomcat, were used as HTTP server and servlet container for Web service, respectively. Perl scripts were used to provide common gateway interface for sequence alignment using ClustalW, and Java applet was used to link Jmol for displaying 3D models. IDBD can be publicly accessed from any Web browser at http://biomarker.cdc.go.kr and http://biomarker.korea.ac.kr.

ACKNOWLEDGEMENTS

We wish to acknowledge the technical support from Mr K.S. Kim from Keewis Co., Seoul, Korea. We extend our thanks to J.H. Lee and H.J. Shin for data collection. We thank Prof. Min Ja Kim for contribution of images of infectious diseases and Prof. Art Cho for his comments on the manuscript. This work was supported by the Korea National Institute of Health, Korea (KNIH). K.J.C. was supported by Post BK21 program of the Ministry of Education, Korea. Funding to pay the Open Access publication charges for this article was provided by the KNIH.

Conflict of interest statement. None declared.

REFERENCES

1. Morens,D.M., Folkers,G.K. and Fauci,A.S. (2004) The challenge of emerging and re-emerging infectious diseases. *Nature*, **430**, 242–249.
2. Li,W., Shi,Z., Yu,M., Ren,W., Smith,C., Epstein,J.H., Wang,H., Crameri,G., Hu,Z. *et al*. (2005) Bats are natural reservoirs of SARS-like coronaviruses. *Science*, **310**, 676–679.
3. Doherty,P.C., Turner,S.J., Webby,R.S. and Thomas,PG. (2006) Influenza and the challenge for immunology. *Nat. Immunol.*, **7**, 449–455.
4. WHO (2004) The world health report 2004 – changing history. *The World Health Report*.
5. Baker,M. (2005) In biomarker we trust? *Nat. Biotechnol.*, **23**, 297–304.
6. Rifai,N., Gillette,M.A. and Carr,S.A. (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.*, **24**, 971–983.
7. Srinivas,P.R., Kramer,B.S. and Srivastava,S. (2001) Trends in biomarker research for cancer detection. *Lancet Oncol.*, **2**, 6.
8. Lee,B.T.K, Song,C.M., Yeo,B.H., Chung,C.W, Chan,Y.L., Lim,T.T, Chua,Y.B., Loh,M.C.S., Ang,B.K. *et al*. (2006) Gastric cancer (biomarkers) knowledgebase (GCBKB): a curated and fully integrated knowledgebase of putative biomarkers related to gastric cancer. *Biomark Insights*, **2**, 135–141.
9. Feng,W., Wu,B., Phan,J., Dale,J., Young,A.N. and Wang,M.D. (2005) An integrated cancer biomarker information system. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **3**, 2851–2854.
10. Lobdell,D.T. and Mendola,P. (2005) Development of a biomarkers database for the national children's study. *Toxicol. Appl. Pharmacol.*, **206**, 269–273.
11. Buonaguro,F.M., Lewis,G.K. and Pelicci,P. (2006) Introducing infectious agents and cancer. *Infect. agents and cancer*, **1**, 1–3.
12. Stephens,S.M., Chen,J.Y., Davidson,M.G., Thomas,S. and Trute,B.M. (2005) Oracle database 10g: a platform for BLAST search and regular expression pattern matching in life sciences. *Nucleic Acids Res.*, **33**, D675–D679.
13. Maglott,D.M., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D266–D268.
14. Laskowski,R.A., Chistyakov,V.C. and Thornton,J.M. (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.*, **33**.
15. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
16. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
17. 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.
18. Stultz,C.M., White,J.V. and Smith,T.F. (1993) Structural analysis based on state-space modeling. *Protein Sci.*, **2**, 305–314.
19. Combet,C., Jambon,M., Deléage,G. and Geourjon,C. (2002) Geno3D an automated protein modelling Web server. *Bioinformatics*, **18**, 213–214.