Published online 10 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D427–D430***
doi:10.1093/nar/gkl880

ApiDB: integrated resources for the apicomplexan bioinformatics resource center

Cristina Aurrecoechea[1], Mark Heiges[1], Haiming Wang[1], Zhiming Wang[2], Steve Fischer[3], Philippa Rhodes[2], John Miller[2], Eileen Kraemer[2], Christian J. Stoeckert Jr.[3], David S. Roos[4] and Jessica C. Kissinger[1,5,*]

[1]Center for Tropical and Emerging Global Diseases and [2]Department of Computer Science, University of Georgia, Athens GA, USA, [3]Department of Genetics, Center for Bioinformatics, 423 Guardian Dr and [4]Penn Genomics Institute, University of Pennsylvania, Philadelphia, PA, USA and [5]Department of Genetics, University of Georgia, Athens GA, USA

Received August 15, 2006; Revised October 10, 2006; Accepted October 11, 2006

ABSTRACT

ApiDB (http://ApiDB.org) represents a unified entry point for the NIH-funded Apicomplexan Bioinformatics Resource Center (BRC) that integrates numerous database resources and multiple data types. The phylum Apicomplexa comprises numerous veterinary and medically important parasitic protozoa including human pathogenic species of the genera *Cryptosporidium, Plasmodium* and *Toxoplasma*. ApiDB serves not only as a database in its own right, but as a single web-based point of entry that unifies access to three major existing individual organism databases (PlasmoDB.org, ToxoDB.org and CryptoDB.org), and integrates these databases with data available from additional sources. Through the ApiDB site, users may pose queries and search all available apicomplexan data and tools, or they may visit individual component organism databases.

INTRODUCTION

The phylum Apicomplexa comprises numerous veterinary and medically important parasitic protozoa including human pathogenic species of the genera *Cryptosporidium, Plasmodium* and *Toxoplasma*. Multiple species of *Plasmodium* are capable of causing malaria in humans, a leading cause of morbidity and mortality in developing countries (1). *Cryptosporidium* causes a severe and chronic diarrheal disease that may be life threatening in immunocompromised patients (2). *Toxoplasma gondii* infections, although typically asymptomatic in healthy individuals, may lead to congenital birth defects and encephalitis in HIV/AIDS patients (3). Human infections with *T.gondii* may be acquired from food or soil contamination and infections with *Cryptosporidium parvum* from soil and water contamination. Due to the potential threat to public health from intentional dispersal into the population, *T.gondii* and *C.parvum* are listed as Category B Biodefense Pathogens by the National Institutes of Health.

The research communities for *Cryptosporidium, Plasmodium* and *Toxoplasma* have benefited from the bioinformatics resources provided by the distinct online genome databases CryptoDB (4), PlasmoDB (5) and ToxoDB (6), respectively (See supplementary material). Because of the phylogenetic relationship of these human pathogens, (all are included in the phylum Apicomplexa, along with the prominent animal pathogens, *Babesia, Theileria* and *Eimeria*) comparative genomic and proteomic studies across these species is critical for expediting discovery of therapeutic targets, increasing understanding of parasite biology and enhancing other areas of research on the biology of these organisms. However, the researcher's ability to perform comparative studies utilizing multiple data sources has been tempered by the difficulty of managing and collating the data from the existing disparate resource databases. Here we describe the online apicomplexan Bioinformatics Resource Center (BRC), ApiDB (http://ApiDB.org), which has been established to provide researchers centralized, integrated access to experimental and computational data, as well as tools to facilitate comparative research.

ApiDB integrates the existing CryptoDB, ToxoDB and PlasmoDB component resources. Database integration is accomplished via a combination of federation and link integration technologies (7). Link integration allows researchers to begin their query with one data source and then follow hypertext links to related information in other data sources. Database federation is achieved by decomposing distributed queries into component queries and executing these queries in the source databases, delivering the results into a uniform format. It leaves the information in its source databases but builds an environment around the databases that makes

*To whom correspondence should be addressed. Tel: +1 706 542 6562; Fax: +1 706 542 3582; Email: jkissing@uga.edu

them all seem part of one large system. In ApiDB 2.0 the federation has been implemented using Oracle DbLink technology. In order to handle heterogeneous data sources in the future we are studying other federation approaches, such as Java Database Connectivity (JDBC) (http://java.sun.com/javase/technologies/database.jsp) and Web Services (WS) (http://www.w3.org/2002/ws/).

ApiDB serves as a web portal for cross-species comparison. Genome data from other apicomplexan parasites are also integrated. In its current release, ApiDB 2.0 offers an initial set of queries that enable gene searches of the three component databases by a variety of criteria such as text keywords, Enzyme Commission (EC) number Gene Ontology (8) assignments, and Pfam (9) terms. In addition, ApiDB offers tools to BLAST (10) all public apicomplexan data, access to the multi-species gene orthology database OrthoMCL DB (11) and access to KEGG (12) metabolic pathway maps with 'painted' comparative highlights of apicomplexan and human enzymes.

FUNCTIONALITY OF CURRENT RELEASE

ApiDB 2.0 was released in April 2006. The datasets available in ApiDB include the component databases (CryptoDB, PlasmoDB and ToxoDB), apicomplexan genomic sequences for other species (*Theileria annulata* and *Theileria parva*) obtained from the NCBI Genbank (13) and GeneDB (14), a collection of clustered apicomplexan ESTs including: *Eimeria, Gregarina, Neospora, Sarcocystis* and *Theileria* called ApiDoTS [a newer version of ApiEST-DB (15)] and unclustered ESTs from the NCBI Genbank division, dbEST.

The ApiDB web interface shares its architecture and 'look and feel' with the component sites (CryptoDB, PlasmoDB and ToxoDB). The user can interact with four areas on the main page: the sidebar, a tools section, a query section and a menu bar (Figure 1A). The sidebar gives the user access to apicomplexan community resources, from our project's most recent news to PubCrawler (16) and external resources, as well as information on the annual ApiDB training workshop (See supplementary material). The tools section provides access to a BLAST search of Apicomplexa genomic, EST and gene model sequences (Figure 1D) and to OrthoMCL DB and KEGG maps with apicomplexan and human enzymes highlighted. The query section provides queries for gene and protein features that span CryptoDB, PlasmoDB and ToxoDB and may include searches of all or a subset of the component species genomes (Figure 1B). Finally, the menu bar appears on every page and gives access to the user's query history and the information on the datasets used in the database.

ApiDB's query architecture provides a set of pages where users can easily execute and manage queries. On the front page, six federated queries are currently available that span the component databases: search genes by gene ID, by annotated keyword in product description (Figure 1B), by Pfam domain, by EC number, by GO term, and by BLAST similarity. Upon query selection, the user is presented with a question page where they can refine their search (Figure 1B). When the query is executed, a summary page offers the number of hits for each organism and the list of genes that meet the requirements (Figure 1C). Hyperlinks connect the user to the gene page in the component sites. Gene pages, acquired from the appropriate component database provide a detailed view of annotation and analysis for the given gene record in the database. For a detailed description of the gene record page we refer to the component databases (4–6).

The query history page, linked in the menu bar, permits users to track their searches and combine them into more complex queries across data types, e.g. find all genes in *Cryptosporidium, Plasmodium* and *Toxoplasma* that have a signal peptide and no transmembrane domains. Summaries of the number of hits for each organism are provided when queries are executed (Figure 1C). As in the component sites, the web interface includes a mechanism to allow users to readily download the sequences and other attributes associated with their query result set (e.g. gene name, product description, coordinates, length) in a versatile tab-delimited file (Figure 1E) that can be viewed in the spreadsheet programs, or, if only sequences are desired, may be downloaded in Fasta format. Examples of inquiries that can be performed on ApiDB are located in the supplementary material.

FUTURE DIRECTIONS

ApiDB will be guided in large part by input from the user communities of ApiDB, the component databases and the objectives of the BRCs. An annual workshop on the usage of apicomplexan database resources is not only a valuable opportunity for users to obtain hands-on instruction, but it also provides a forum for feedback used to further drive development of this site. As the autonomous component databases evolve with new data and features, ApiDB will respond to integrate these elements as appropriate. As data from phylogenetically related species becomes increasingly available, e.g. ciliates and *Perkinsus*, and perhaps, some day a dinoflagellate, orthologous genes will be determined and links will be provided, via orthology to these resources when possible.

The ApiDB website currently excludes queries of datasets not found in all three component databases. For example, PlasmoDB contains microarray-based gene expression data whereas CryptoDB and ToxoDB presently do not. We will investigate permitting the querying of a subset of the component databases through ApiDB. Feature enhancements that will be available by early 2007 include persistent history to allow users to save their search results and corresponding results and a 'sort' tool to allow users to sort the results they obtain from individual queries (Figure 1C) by species, gene ID or feature description.

Combinations of existing tools, pre-formed queries and query history can be leveraged in powerful ways to mine the apicomplexan data but do not lend themselves well to high-throughput explorations. To facilitate large-scale database utilization, we will be providing programmatic access to our facilities through the use of standard web service technologies. Modular web service tasks will serve as building blocks for creating a user-defined workflow. As a single example, linking a series of tasks would enable a researcher with a collection of putative gene regulatory

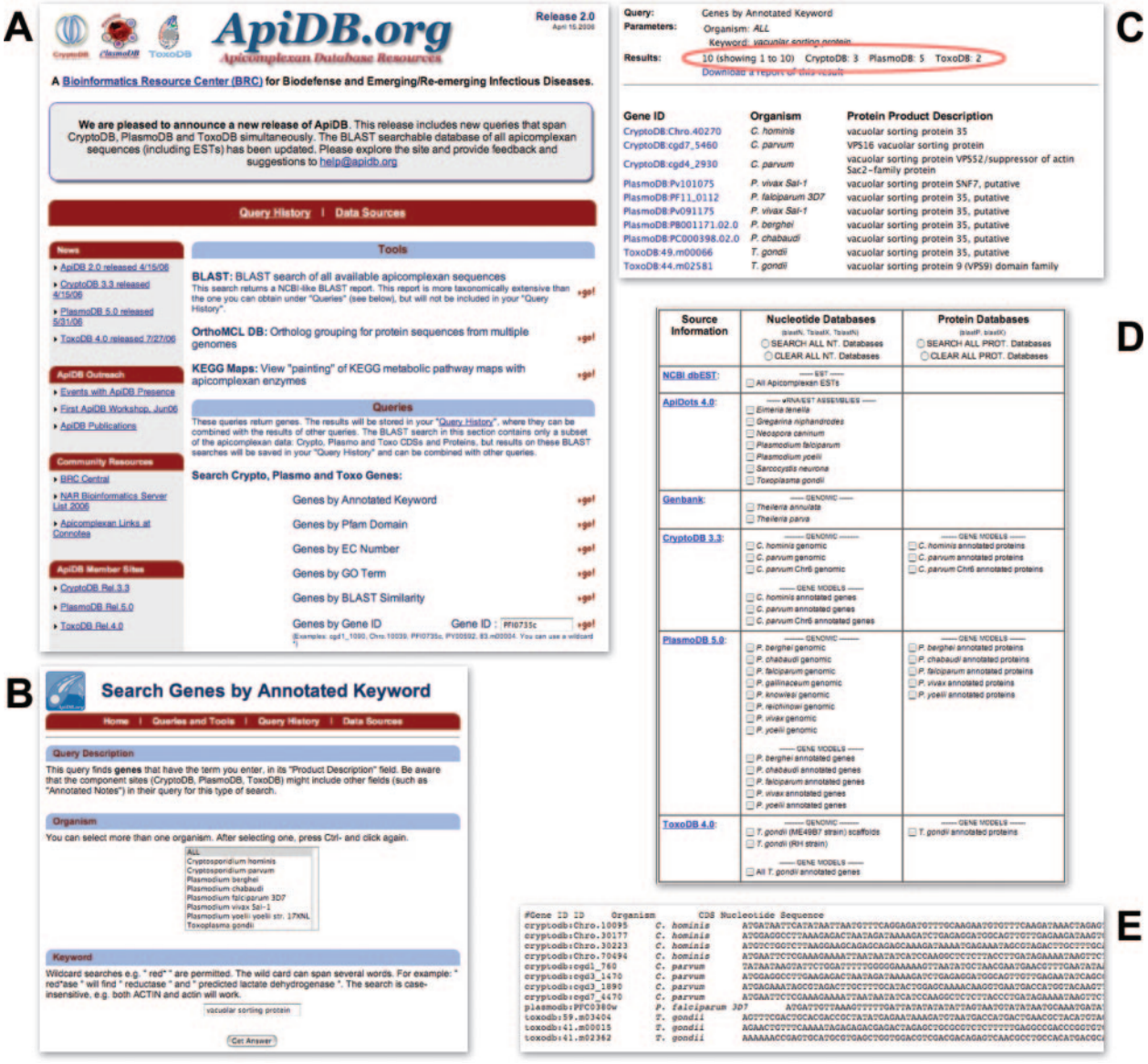

Figure 1. Database Functionality (**A**) Searches are initiated via tools or queries provided on the web site's front page. (**B**) Query forms allow user-defined refinement for the queries, e.g. which species would you like to search? Descriptions and help are provided for the queries. (**C**) Query searches return a table with results from the component databases sites. The result summary states the number of records found at each site (circled). Individual results are linked to a detailed record page at its component database site. (**D**) The BLAST tool integrates searches of all publicly available apicomplexan genomic data, including species not hosted by a component database. (**E**) Query results can be downloaded in a customized tab-delimited file.

motifs to perform a bulk series of BLAST searches of chromosomal sequences, extract hit coordinates from the report, then query for gene models downstream of each hit. As our web services capabilities expand, we will use them to permit workflows that include ClustalW (17) for multiple sequence alignments. Web services may also be used to collaborate with additional databases, including other BRCs (http://BRC-central.org), to offer users integrated access to additional pathogen data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The ApiDB project is co-administered by David S. Roos, Christian J. Stoeckert Jr and Jessica C. Kissinger. We thank our database development collaborators: Aaron J. Mackey, Bindu Gajria, Thomas Gan, Jerric Gao, John Iodice and the entire development team for the component database sites. This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN26620040 0037C. Funding to pay the Open Access publication charges for this article was provided by National Institute of Allergy and Infectious 215 Diseases, National Institutes of Health,

Department of Health and Human Services, under Contract No. HHSN26620040 0037C.

Conflict of interest statement. None declared.

REFERENCES

1. Lopéz-Antuñano,F.J. and Schmunis,G.A. (1993) *Parasitic Protozoa*. Academic Press Inc., San Deigo, CA, Vol. 5, pp. 135–265.
2. Fayer,R. (1997) *Cryptosporidium and Cryptosporidiosis*. CRC Press Inc., Boca Raton, FL.
3. Dubey,J.P. and Beattie,C.P. (1988) *Toxoplasmosis of Animals and Man*. CRC Press Inc., Boca Raton, FL.
4. Heiges,M., Wang,H., Robinson,E., Aurrecoechea,C., Gao,X., Kaluskar,N., Rhodes,P., Wang,S., He,C.-Z., Su,Y. *et al*. (2006) CryptoDB: a *Cryptosporidium* bioinformatics resource update. *Nucleic Acids Res*., **34**, D419–D422.
5. Bahl,A., Brunk,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C., Labo,P. *et al*. (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res*., **31**, 212–215.
6. Kissinger,J.C., Gajria,B., Li,L., Paulsen,I.T. and Roos,D.S. (2003) ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res*., **31**, 234–236.
7. Stein,L.D. (2003) Integrating biological databases. *Nature Rev. Genet*., **4**, 337–345.
8. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
9. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res*., **28**, 263–266.
10. 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.
11. Chen,F., Mackey,A.J., Stoeckert,C.J.,Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*., **34**, D363–D368.
12. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.
13. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res*., **31**, 23–27.
14. Hertz-Fowler,C., Peacock,C.S., Wood,V., Aslett,M., Kerhornou,A., Mooney,P., Tivey,A., Berriman,M., Hall,N., Rutherford,K. *et al*. (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res*., **32**, D339–D343.
15. Li,L., Crabtree,J., Fischer,S., Pinney,D., Stoeckert,C.J., Sibley,L.D. and Roos,D.S. (2004) ApiEST-DB: analyzing clustered EST data of the apicomplexan parasites. *Nucleic Acids Res*., **32**, D326–D328.
16. Hokamp,K. and Wolfe,K.H. (2004) PubCrawler: keeping up comfortably with PubMed and GenBank. *Nucleic Acids Res*., **32**, W16–W19.
17. Jeanmougin,F., Thompson,J.D., Gouy,M., Higgins,D.G. and Gibson,T.J. (1998) Multiple sequence alignment with ClustalX. *Trends Biochem. Sci*., **23**, 403–405.