PlasmoDB: An integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data

The Plasmodium Genome Database Collaborative*

Department of Biology, University of Pennsylvania, 415 South University Avenue, Philadelphia, PA 19104-6018, USA

Received October 2, 2000; Revised and Accepted November 1, 2000

ABSTRACT

The *Plasmodium falciparum* Genome Database (http://PlasmoDB.org) integrates sequence information, automated analyses and annotation data emerging from the *P.falciparum* genome sequencing consortium. To date, raw sequence coverage is available for >90% of the genome, and two chromosomes have been finished and annotated. Data in PlasmoDB are organized by chromosome (1–14), and can be accessed using a variety of tools for graphical and text-based browsing or downloaded in various file formats. The GUS (Genomics Unified Schema) implementation of PlasmoDB provides a multi-species genomic relational database, incorporating data from human and mouse, as well as *P.falciparum*. The relational schema uses a highly structured format to accommodate diverse data sets related to genomic sequence and gene expression. Tools have been designed to facilitate complex biological queries, including many that are specific to *Plasmodium* parasites and malaria as a disease. Additional projects seek to integrate genomic information with the rich data sets now becoming available for RNA transcription, protein expression, metabolic pathways, genetic and physical mapping, antigenic and population diversity, and phylogenetic relationships with other apicomplexan parasites. The overall goal of PlasmoDB is to facilitate Internet- and CD-ROM-based access to both finished and unfinished sequence information by the global malaria research community.

BACKGROUND

Plasmodium falciparum is the causative agent of cerebral malaria, which afflicts much of the world's population and is responsible for the death of more than one million people each year, mostly children in sub-Saharan Africa. Three other species of *Plasmodium* parasites also cause human malaria, and other members of the phylum Apicomplexa are also serious human and veterinary pathogens. The *P.falciparum* genome project was initiated in 1996 by an international consortium, with support from private and government agencies in both the UK and the US (1). The success of the *Plasmodium* genome project will ultimately be determined by how rapidly and effectively the information it produces is utilized by the research community to advance our understanding of malaria. In particular, effective dissemination of genomic information should accelerate the development of new therapeutics and vaccines. This article describes bioinformatics resources designed to help move information from the genome project and related research into the hands of malaria researchers worldwide.

The *Plasmodium* genome database project seeks to:

- Incorporate data and annotation emerging from the *P.falciparum* genome sequencing centers into a searchable resource, providing a single, robust point of access and a user-friendly interface, with extensive links to relevant resources and databases under development elsewhere.
- Provide additional views of finished and unfinished sequence data based on automated analysis.
- Integrate genomic data with existing and future experimental results related to genome mapping, gene expression, protein function, protein–protein interactions, antigenic diversity, metabolic pathways, etc, in both graphical and text-based formats.
- Establish a database that can incorporate sequence information for other *Plasmodium* species and related apicomplexan parasites.
- Facilitate expert curation of genomic and other related data.
- Provide web-based access, email servers (including an online 'help desk') and CD-ROM distribution, in order to ensure full access to the *Plasmodium* genome and related resources for all interested parties regardless of geographic location.

SCOPE

Species

PlasmoDB was designed to accommodate data from numerous *Plasmodium* and/or relevant apicomplexan species, focusing on the malarial parasite *P.falciparum* (strain 3D7; with some

*Correspondence should be addressed to: David S. Roos, Department of Biology, University of Pennsylvania, 415 South University Avenue, Philadelphia, PA 19104-6018, USA. Tel: +1 215 898 2118; Fax: +1 215 898 8780; Email: droos@sas.upenn.edu

Current members of the *Plasmodium* Genome Database Collaborative include: J. C. Kissinger, A. J. Milgram, M. J. Fraunholz, and D. S. Roos (Department of Biology, University of Pennsylvania); B. P. Brunk, J. Crabtree, S. J. Diskin, J. Schug, G. C. Overton, and C. J. Stoeckert Jr (Center for Bioinformatics, University of Pennsylvania); R. L. Coppel and R. L. Huestis (Deparment of Microbiology, Momash University).

additional data from strains B8 and FCR3). All *Plasmodium* data presently available through PlasmoDB relate to *P.falciparum.*

Data

PlasmoDB can accommodate and integrate diverse types of data, including, but not limited to: genomic DNA sequence (finished and unfinished); microsatellite and physical mapping information; predicted translations, protein features and motifs, GO classifications; BLAST results; EST, STS or GSS sequences; expression data from microarray and SAGE analyses; and newly emerging proteomics data. At present, finished and unfinished genomic sequence data and BLAST results are available. For chromosomes that have been completed and annotated, predicted translations of identified coding regions and various protein features are also available.

CONTENT OF CURRENT RELEASE

The current release of PlasmoDB is divided into four main areas (discussed below) providing different views and approaches to data analysis: Genomics Unified Schema (GUS), GenePlot, Data Mining/Download and Links. The tools provided in each of these areas overlap to a degree, as demanded by the services that they provide to distinct user communities. The database incorporates all released genomic DNA sequence data (finished and unfinished), integrated with annotation provided by the genome sequencing centers and additional automated analyses from various sources. The current (October 25, 2000) release of PlasmoDB contains two complete chromosome (2,3) and 3523 contig sequences representing all 14 chromosomes and totaling 27 606 283 nt.

GUS (Genomics Unified Schema)

General format. The GUS section of PlasmoDB is a relational database containing draft and finished genomic sequence, ESTs, genes and gene predictions, along with a variety of automated analyses such as predicted GO function and similarities to Pfam and ProDom domains. The web interface permits a large collection of sophisticated queries and views of the data. GUS is a rich relational schema for genomic data developed by the Computational Biology and Informatics Laboratory at the University of Pennsylvania, and contains ~190 tables (4). The schema is not species-specific, permitting storage of (and convenient access to) genomic data from human, mouse and other species in the same database, facilitating integration with *P.falciparum* data as required. GUS is organized around the ‘central dogma’ of biology: DNA is transcribed and spliced to generate mRNA, which is translated into protein. This principle allows key biological concepts to be integrated into an organizational framework for queries and further analysis. GUS employs the GO gene ontology framework (5) and other structured vocabularies to describe gene function, species, tissue and developmental stage, providing a consistent annotation for developing effective queries. Evidence tables link manual conclusions and computational predictions to the raw data on which the decisions were based. GUS also tracks ownership, protection and the source of data on a row-by-row basis, permitting the tracking of data by source and project.

Resources and tools. The GUS section of PlasmoDB provides graphical and tabular views of genes and predictions by chromosome, and detailed information about individual (predicted) genes. Graphical views of genomic sequence are available in both GIF and Java applet form. Parameterized queries select genes or gene predictions, e.g. by location, exon size, presence of Pfam or ProDom domains, predicted GO function, presence of signal peptides or transmembrane domains, etc. The interface keeps track of queries executed in each session, and these results can be recalled via a query history page. The web interface also permits Boolean combinations of query results, allowing users to combine selected queries by intersection or union. Complex Boolean queries can be built before execution, using a form that determines which attributes will be queried, and permitting the user to enter subquery parameters (e.g. chromosomal location or Pfam domain). Query outline can be bookmarked for later use with the same or different parameters.

GenePlot

General format. The *P.falciparum* GenePlot database is a stand-alone platform-independent compilation of all available finished and unfinished DNA sequence for the *P.falciparum* genome. The data is accessible online or via CD-ROM using any JavaScript-enabled web browser (Internet Explorer or Netscape version 4.0 or higher are recommended). Particular attention is devoted to ease of use and distribution, recognizing the worldwide interest in malaria, including researchers without convenient access to high-speed Internet connections or reliable, inexpensive telephone lines. Accordingly, GenePlot is designed to facilitate search and retrieval of DNA sequence information for further analysis using locally available tools.

Resources and tools. Both finished and unfinished contig sequence data are organized by chromosome. Finished, annotated sequences (such as chromosomes 2 and 3) provide links to feature indices (gene names) contained on each sequence or contig, and graphical displays of the annotated features. Images display information about the gene products, including mouse-over identification of features. DNA and predicted protein sequences corresponding to these features can be retrieved dynamically, using links from the image, or through the indices. Complete or partial DNA sequence and predicted translations of genes, introns, annotated features or unannotated regions specified by the user can also be dynamically retrieved using tools provided on a separate sequence-retrieval page and available on every page of the annotated sequence information.

Data mining tools/data download

General format. This section of PlasmoDB provides an additional set of web-based tools emphasizing data mining from unfinished and unannotated sequence. First pass annotation of the entire *P.falciparum* reference strain is anticipated to be available early in 2001. Additional sequencing projects providing low-pass coverage of other *Plasmodium* species, related parasites and additional strains of *P.falciparum* (including diverse field isolates) are expected to yield a continuous stream of unfinished data, mandating the availability of tools permitting rapid, automated analysis of such sequences based on user-defined queries.

These tools are designed so that results are displayed immediately, with links for the retrieval of original sequences if desired; contig sequence names have been designed to permit trace-back to the sequencing center if necessary. A web browser of version 4.0 or higher is required.

Resources and tools.

- Download the current and past versions of contig sequence data. All available genomic DNA sequence data from the various sequencing centers involved in the *P.falciparum* genome project is available for retrieval, on a whole genome or chromosome-specific basis. Multiple sequence formats are provided (Fasta, GenBank, EMBL). Files of all open reading frames (ORFs) >20, 50, 75 or 100 amino acids are also available for the entire genome. Where available, the nucleotide sequences and predicted coding sequence for processed RNAs can also be downloaded.
- BLAST, BLASTn, tBLASTn and tBLASTx searches against the most recent releases of finished and unfinished sequence contigs for the entire genome. BLAST results provide a hot-linked graphical interface, and links to identified sequences, facilitating further examination and analysis by the user.
- Text based queries. All *P.falciparum* contigs have been analyzed using BLASTx searches against the non-redundant GenBank protein database, post-processed using MSPcrunch (6) and compiled as a searchable index based on key words found in the description lines of BLAST results. Text-based user queries (using Boolean operations if desired) return all significant BLASTx hits (*E*-values of $\leq 10^{-3}$) between *P.falciparum* contig sequence queries and GenBank subjects containing the requested key word(s). Results are displayed along with the alignment, links back to the original *P.falciparum* contig sequence and a graphical display of all other GenBank hits encountered in the same contig.
- User-defined motif search. This tool allows researchers to define their own protein motif and then search a crude translation of all available *P.falciparum* sequence data to see if the motif is present. Links back to the contig sequence are provided.

Links

Listing of links for malaria researchers to a variety of useful resources, including the centers responsible for generating *P.falciparum* sequence data and annotation:

- Sanger Centre: http://www.sanger.ac.uk/Projects/P_falciparum/;
- TIGR/NMRC: http://www.tigr.org/tdb/edb/pfdb/pfdb.html;
- Stanford University: http://sequence-www.stanford.edu/group/malaria/index.html;

and many malaria-specific resources including:

- Malaria Foundation: http://www.malaria.org/;
- Malaria Research and Reference Reagent Resource Center (MR4): http://www.malaria.mr4.org/mr4pages/index.html;
- NCBI Malaria Genetics and Genomics page: http://www.ncbi.nlm.nih.gov/Malaria/;
- Malaria Database: http://www.wehi.edu.au/MalDB-www/who.html;
- *Plasmodium* transcript reconstruction project: http://www.sanbi.ac.za/malaria-genesearch/;
- Malaria Antigen Database: http://ben.vub.ac.be/malaria/mad.html;
- Rodent Malaria Index: http://www.ncbi.nlm.nih.gov/Malaria/Rodent/index.html;
- Mosquito genomics WWW server: http://klab.agsci.colostate.edu/;

and other relevant links too numerous to list.

FUTURE PLANS

PlasmoDB is scheduled to be updated every two months. Distribution of GenePlot in CD format is coordinated with the World Health Organization's CD release program (approximately twice yearly; for further information, contact ross.coppel@med.monash.edu.au).

Tools and features currently under development and scheduled for release within the next year include:

- Incorporation of gene predictions using multiple complementary gene-finding algorithms (7,8), and other features identified via automated annotation (putative signal sequences, transmembrane domains and organellar targeting signals; nucleotide and codon bias; etc.).
- Integration of *P.falciparum* genomic information with genetic (microsatellite) (http://www.ncbi.nlm.nih.gov/Malaria/markers_maps.html) (9), optical (10) and other mapping data into a common graphical map.
- Addition of dynamic translation tools, motif identification and BLAST results to the CD version of GenePlot.
- Incorporation of additional data types into the GUS schema, especially expression data from *P.falciparum* and new sequence data from additional species of *Plasmodium* and related parasites (GUS is capable of cross-species comparisons).

The *Plasmodium* Genome Consortium (including funding agencies, sequencing centers, bioinformatics groups and researchers) has expressed a strong commitment to supporting the growth and development of informatics efforts, to ensure maximum gains from past and future investment in sequencing and other genomics projects.

ACCESS AND REFERENCING

PlasmoDB is located at http://PlasmoDB.org and GenePlot is available via the Internet at PlasmoDB, or on a stand-alone CD-ROM.

To obtain a GenePlot/WHO malaria CD-ROM contact Ross Coppel, Department of Microbiology, Monash University, Clayton, Victoria 3800, Australia. Email: ross.coppel@med.monash.edu.au. For additional help, suggestions, or to report errors and/or submit data, contact PlasmoDB c/o corresponding author or PlasmoDB@pcbi.upenn.edu.

We suggest that PlasmoDB be referenced as follows: The *Plasmodium* Genome Consortium (2001) PlasmoDB: an integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. *Nucleic Acids Res.,* **29**, 66–69.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The *Plasmodium* Genome Database, PlasmoDB, is supported by a grant from the Burroughs Wellcome Fund. Additional computational support was provided by Rahul Dave, Haralabos Tikos and the Eniac2000 project (supported in part through equipment donations from Intel).

The scientists and funding agencies comprising the international Malaria Genome Sequencing Consortium agree that sequence data of the genome of *Plasmodium falciparum* (3D7) should be made public before publication of the complete sequence. Sharen Bowman, Neil Hall, Bart Barrell and their colleagues at the Sanger Center (UK) are sequencing chromosomes 1, 3, 4, 5, 6, 7, 8, 9 and 13. Work at the Sanger Center is supported by the Wellcome Trust. The members of the USA consortium composed of Malcolm Gardner, Leda Cummings, Claire Fraser and their colleagues at The Institute for Genome Research (TIGR) and Dan Carucci, Steven Hoffman and their colleagues at the Navy Military Research Center (NMRC) are sequencing chromosomes 2, 10, 11 and 14. Work at TIGR/NMRC is supported by the NIAID/NIH, the Burroughs Wellcome Fund, and the Department of Defense (USA). Richard Hyman, Eula Fung, Ron Davis and their colleagues at the Stanford Genome Technology Center (USA) are sequencing chromosome 12. The work at Stanford University is supported by the Burroughs Wellcome Fund.

REFERENCES

1. Fletcher,C. (1998) The *Plasmodium falciparum* Genome Project. *Parasitology Today*, **14**, 342–344.
2. Gardner,M.J., Tettelin,H., Carucci,D.J., Cummings,L.M., Aravind,L., Koonin,E.V., Shallom,S., Mason,T., Yu,K., Fujii,C. *et al.* (1998) Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science*, **282**, 1126–1132.
3. Bowman,S., Lawson,D., Basham,D., Brown,D., Chillingworth,T., Churcher,C.M., Craig,A., Davies,R.M., Devlin K., Feltwell,T. *et al.* (1999) The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature*, **400**, 532–538.
4. Davidson,S., Crabtree,J., Brunk,B., Schug,J., Tannen,V., Overton,C. and Stoeckert,C. (2000) K2 and GUS: Experiments in integrated access to genomics data sources. *IBM Systems J.*, **40**, in press.
5. 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.
6. Sonnhammer,E.L.L. and Durbin,R. (1994) A workbench for large-scale homology analysis. *Comput. Appl. Biosci.*, **10**, 301–307.
7. Pertea,M., Salzberg,S.L. and Gardner,M.J. (2000) Bioinformatics: Finding genes in *Plasmodium falciparum*. *Nature*, **403**, 34.
8. Lawson,D., Bowman,S. and Barrell,B. (2000) Bioinformatics: Finding genes in *Plasmodium falciparum*. *Nature*, **403**, 34.
9. Su,X., Ferdig,M., Huang,Y., Huynh,C.Q., Liu,A., You,J., Wootton,J.C. and Wellems,T.E. (1999) A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science*, **286**, 1351–1353.
10. Lai,Z., Jing,J., Aston,C., Clarke,V., Apodaca,J., Dimalanta,E.T., Carucci,D.J., Gardner,M.J., Mishra,B., Anantharaman,T.S. *et al.* (1999) A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genet.*, **23**, 309–313.