commentary

The *Plasmodium* genome database

Designing and mining a eukaryotic genomics resource.

The Plasmodium Genome Database Collaborative

As reported elsewhere in this issue (M. J. Gardner *et al. Nature* **419**, 498–511; 2002), a reference genome sequence for the human malaria parasite *Plasmodium falciparum* is now complete. But how are researchers to access *P. falciparum* genome sequence data, integrate this resource with other relevant data sets, and exploit the resulting information for functional studies, including identification of novel drug targets and candidate vaccine antigens?

The *Plasmodium* genome database (PlasmoDB, see http://PlasmoDB.org) contains information from multiple sources, including DNA sequence data and curated annotations, automated gene model predictions, predicted proteins and protein motifs, cross-species comparisons, optical and genetic mapping data, information on population polymorphisms, expression data generated by a variety of complementary strategies, and proteomics data. Integrating this information at a single site provides 'one-stop shopping' for genomics-scale data sets related to malaria parasites.

The use of a relational database architecture enables users to ask complex questions. For example, immunologists trying to develop an anti-malaria vaccine might wish to identify potential immunodominant surface antigens. Drug developers might wish to identify enzymes expressed in bloodstream parasites that differ significantly from their human counterparts. Researchers interested in antigenic variation and how the parasite adheres to cells (a cause of malaria pathogenesis) might wish to identify all gene families in the parasite genome; those interested in genome organization might be interested in the chromosomal location of these proteins; evolutionary biologists might wish to examine all genes for which clear orthologues are known from a range of species; and so on.

Universal access

It has taken six years to complete the *P. falciparum* genome sequence. In the meantime, interim data were periodically released by the three sequencing centres involved in this project, to advance research on basic malaria biology, drug and vaccine development. PlasmoDB was developed to make this information available to the research community, notwithstanding the challenges posed by unfinished sequence data. This web-accessible database provides access to the entire genome sequence of the 3D7 reference strain of *P. falciparum*, together with computationally predicted and manually curated genes and gene models, protein feature predictions and functional annotation.

PlasmoDB went live in June 2000 — more than two years before today's formal completion of the *P. falciparum* reference sequence. The website receives several thousand hits each day from more than 100 countries, numbers that are certain to rise significantly with the release of the complete genome sequence. The result can be measured in the scores, possibly hundreds, of publications that have resulted, and in new targets now being assessed for drug and vaccine development.

Malaria biologists are a more diverse and dispersed community than those who study fruitfly or yeast genomes. They encompass field scientists in Cameroon, epidemiologists in Papua New Guinea, pharmaceutical developers in India, molecular geneticists in Brazil, and so on. Because many malaria researchers lack reliable high-speed Internet access, a platform-independent CD-ROM (to be distributed with *Nature* in a few weeks' time) has been developed to provide universal free access to the complete genome sequence and annotations currently available for this malaria parasite. More than a series of 'flat-file' images, *P. falciparum* GenePlot is a true database, providing a graphical user interface for browsing, querying, downloading and manipulating the genome and annotations on a desktop computer without web access.

It has been a stimulating challenge to see how many commonly asked questions can be accommodated in the CD-ROM format. For example, while local implementation of BLAST searches requires substantial memory and computational speed (and GenBank is too large to include on a single CD), GenePlot can be asked to find and retrieve all predicted proteins with similarity to proteases, based on text indices derived from precomputed BLAST comparisons of the entire *P. falciparum* genome against all of GenBank.

The initial motivation behind the GenePlot CD was to make the genome accessible to malaria biologists with limited Internet connectivity, but this format has also proved enormously popular with well-connected users. Having the data literally 'in hand' provides scientists everywhere with a sense of ownership and involvement in the *Plasmodium* genome project, expediting the pace of research and discovery related to malaria parasites and the devastating diseases they cause.

Unfinished business

In most genomics projects, initial mapping studies (desirable even with the advent of whole-genome shotgun sequencing) are followed by a random sequencing phase, then by a phase focusing on closure of remaining gaps to produce a 'finished' sequence (which may still contain numerous gaps, depending on complexity and size of the genome, time, patience and funding). Annotation is conducted to various levels of depth. Database development makes the information accessible to the user community. Finally, functional studies (transcript profiling, proteomic studies, genome-scale knockouts, and so on) become possible once the complete, annotated sequence is available to end-users, typically via the Internet.

There are good reasons for this sequential strategy. Gap closure is expensive, and so makes little sense while random sequencing may still yield useful information. Manual annotation of assembled sequences is also laborious, and is best deferred until the genome sequence is complete. For large complex eukaryotic genomes, years may pass between the initial sequencing and the availability of this information in practical form for researchers in the lab. Such delays cause considerable frustration, as virtually all individual genes can be identified long before assembly of a finished genome.

Problems associated with unfinished data, and the accompanying need for user education regarding the interpretation of these results, provided the first challenge for PlasmoDB. Specific information missing in incomplete data

The CD-ROM containing *P. falciparum* GenePlot and other malaria-related resources, including *Nature*'s malaria Insight of 7 February 2002 and the papers reported elsewhere in this issue, will be provided to all *Nature* subscribers in a few weeks' time. It can also be obtained from helpcd@plasmodb.org or the Malaria Research and Reference Reagent Resource Center (MR4) by an e-mail request to malaria@atcc.org, with "Nature malaria CD-ROM" in the subject line. A full postal address must be included in the body of the message.

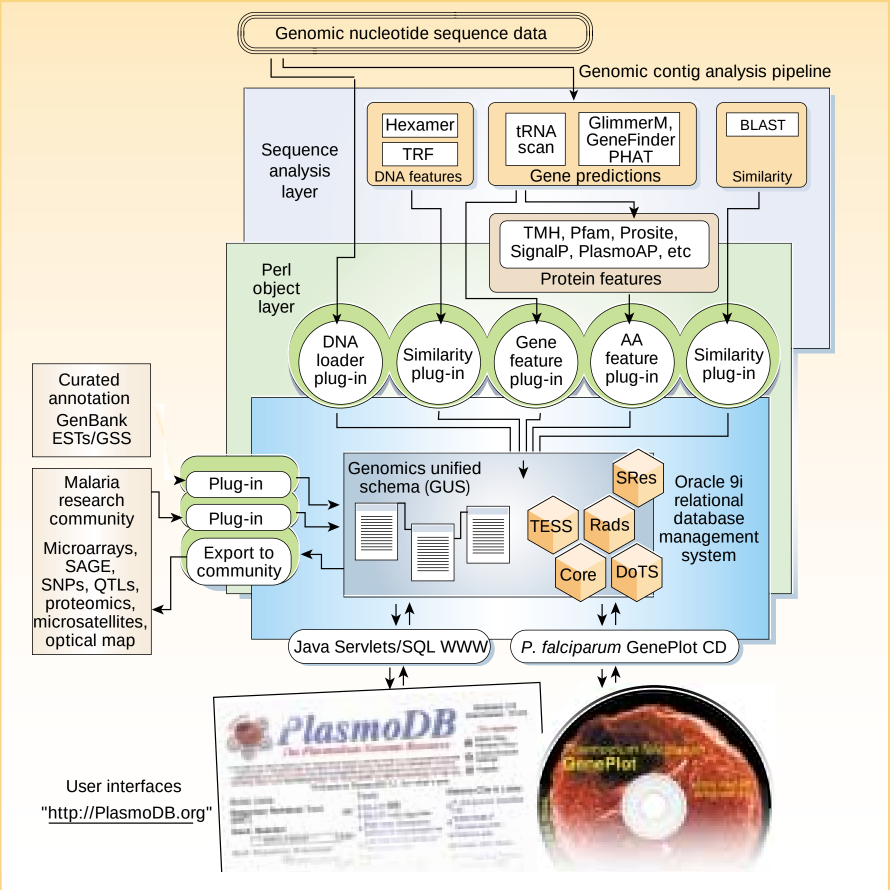
sets limits confidence that a particular gene is absent from the organism. Contaminating sequences from cloning vectors and host cells may be present. Redundancy in the dataset attributable to incomplete or inaccurate assemblies poses a further problem, particularly for the A/T-rich *P. falciparum* genome. In PlasmoDB, possible redundancy or inaccurate assembly was identified by high-stringency comparisons of each sequence with the entire genome; and comparison of DNA sequences with optical and genetic maps. The importance of these tools for *P. falciparum* declines as the genome project approaches completion, but they remain valuable for new projects, such as the other *Plasmodium* species being sequenced.

Unfinished sequence data also pose challenges for gene identification and analysis, as the constantly changing nature of this information makes time-consuming manual annotation impossible. Comparisons with GenBank, computational gene-finding algorithms and protein feature analyses are feasible (Box 1), but generate a bewildering range of predictions: which of four competing gene predictions is most likely to be correct? Which of 60 sequences exhibiting similarity to cathepsin is really a protease? Automated analysis can help to provide provisional assignments early, before manual curation of the finished sequence. Even after first-pass annotation, these analyses can help to suggest alternative possibilities whenever new experimental information suggests inaccuracies in the curated annotation.

Integrative 'omics'

Many disciplines accommodate large data sets (MRI imaging, weather forecasting, ecological and econometric modelling, and so on), but this is a relatively new problem for molecular and cell biologists. How to collect the deluge of data engulfing us from genomics, transcriptomics, proteomics, glycomics, pharmacogenomics, vaccinomics, and even more hideously named approaches? What kind of tools will be required to analyse — and to integrate — these massive 'omics'-scale data sets? How can we use all this information to treat malaria?

PlasmoDB is based on a relational database architecture (GUS; Box 1), built around biologically relevant relationships following the central dogma of biology: 'gene to messenger RNA to protein'. Parallel views for other organisms (including other *Plasmodium* species) enable phylogenetic comparisons. Because all this information is in a single database, queries can combine searches for particular genes of interest with RNA and protein expression analysis, studies on population genetic polymorphisms, and cross-species comparison. One can envisage the incorporation of other data types, such as publication records, clinical outcome data, genomic information from the mosquito vector *Anopheles gambiae*, protein structural information


Genomic nucleotide sequence data
Genomic contig analysis pipeline
Sequence analysis layer
Hexamer
TRF
DNA features
tRNA scan
GlimmerM, GeneFinder PHAT
Gene predictions
BLAST
Similarity
TMH, Pfam, Prosite, SignalP, PlasmoAP, etc
Protein features
Perl object layer
DNA loader plug-in
Similarity plug-in
Gene feature plug-in
AA feature plug-in
Similarity plug-in
Curated annotation GenBank ESTs/GSS
Malaria research community
Microarrays, SAGE, SNPs, QTLs, proteomics, microsatellites, optical map
Plug-in
Plug-in
Export to community
Genomics unified schema (GUS)
SRes
TESS
Rads
Core
DoTS
Oracle 9i relational database management system
Java Servlets/SQL WWW
P. falciparum GenePlot CD
User interfaces
"http://PlasmoDB.org"
PlasmoDB
GenePlot

Box 1: The architecture of PlasmoDB

PlasmoDB is not itself a database, but a web interface that uses an underlying relational database (GUS, for genomics unified schema), which stores and integrates nucleotide sequences, annotation, information on gene expression and regulation, controlled vocabularies/ontologies, and evidence for these annotations. GUS is organism-independent and also contains the human and mouse genomes (http://www.allgenes.org). The schema, associated code and project-independent data are at http://www.gusdb.org.

Primary *P. falciparum* sequence data are subjected to automated analyses (sequence analysis layer), including the identification of motifs and simple repeats; comparison against the entire genome to identify gene families, repetitive elements and redundancy; searching for intron/exon structure, using several algorithms trained on experimentally validated *P. falciparum* sequences; conceptual gene translation and identification of potential protein motifs; and comparisons with the non-redundant GenBank/EMBL database (results retained in a text-queryable index). Genomic contig sequences are aligned to optical restriction maps and microsatellite linkage groups using hidden Markov models for fragment length and ePCR.

The GUS schema uses views that are used in an object layer for parent–child relationships. To facilitate data loading, Perl was used to create a 'thin' object layer in which each relational table is treated as an object. GUS is partitioned into distinct name spaces. Core contains workflow tables, tracking how each row in the database is populated (data provenance). Sres (shared resources) contains controlled vocabularies and ontologies, such as taxonomy, anatomy and disease tables. TESS captures descriptions (grammar representations) for genetic regulatory regions (not currently implemented for PlasmoDB). DoTS houses sequence and sequence annotation. Any sequence span can have multiple features mapped to it, and gene predictions can be associated with multiple transcripts and proteins. Each predicted or experimentally determined transcript may itself have multiple features and similarities, as can each protein entry. RAD handles data from high-throughput technologies for studying gene or protein expression. RAD currently accommodates expression data from EST projects, SAGE (serial analysis of gene expression) studies , cDNA and oligonucleotide glass slide microarrays, and Affymetrix chips, and is extensible to accommodate information from other platforms. Sample information, together with other experimental descriptions, can be entered directly into the database via web-based forms.

The RAD schema is compliant with MIAME guidelines (http://www.mged.org/). A microarray gene expression (MAGE) object model and XML-based language have been developed for data exchange, and importers and exporters are being built for RAD to MAGE-ML.

from high-throughput crystallography studies, and chemical compound libraries.

The power of database queries

PlasmoDB provides graphic and text-based views of all available *Plasmodium* genomic sequences, curated annotations, and tools for retrieval of these data. But the sheer wealth of information can make browsing difficult, so the database allows the user to define custom views. For all their visual appeal, however, static, precomputed views are inherently local, and so fail to answer many genomic-scale questions that arise in the laboratory.

The relational database underlying PlasmoDB permits questions to be asked that integrate diverse data types, as illustrated by questions relating to drug and vaccine development (Table 1). For example, a medicinal chemist might be interested in *P. falciparum* dihydrofolate reductase (DHFR), the target of the drug pyrimethamine used in common antimalarial agents. The gene encoding this enzyme can be identified by text searches of curated annotation using the enzyme name (or EC number) as a key word; by text searches against BLAST results, by using conserved protein sequence signatures for a motif search; by BLAST similarity to DHFR sequences from other species; or by searches based on protein structural predictions. Degenerate searches are also possible, such as searching for all proteases. The results returned would undoubtedly contain false positives, but these can be weeded out by scientists familiar with protease characteristics. Candidate cyto-skeletal proteins can be identified by similar strategies, or by searching for protein structural predictions. Such searches can then be refined, for example by identifying sequences conserved in multiple malaria parasites, or those that are sufficiently distinct from human orthologues to provide a basis for selective inhibition.

Information on metabolic pathways and/or subcellular localization can also be used to inform database queries. For example, PlasmoDB enables the identification of proteins likely to be associated with the apicoplast — a distinctive organelle that has received considerable attention as a candidate drug target — on the basis of curated annotation, exploiting the structured gene ontology (GO) vocabulary. Alternatively, the origins of this organelle by horizontal transfer of an algal chloroplast can be exploited as the basis for a text search for genes exhibiting sequence similarity to plastid, chloroplast or plant genes. Phylogenetic comparison with plant species is not currently supported in PlasmoDB, but all nucleotide and predicted protein sequences can be downloaded by users for local analysis.

Combining gene and protein predictions with the results from RNA and/or protein expression analysis enables enzymes being considered for antimalarial drug development to be filtered, removing any proteins not expressed in blood-stage parasites. Integrating these data with others, such as functional studies, publications, or small molecule databases, allows further refinement.

Table 1 **Querying *Plasmodium* genome data**

User query	*Computational strategy/approach*
Drug development	
Dihydrofolate reductase	Text*, EC*, motif* and BLAST* searches
Proteases	Text* and motif* searches; GO function assignment*
Cytoskeletal genes conserved in multiple *Plasmodium* species	GO cellular component assignment*; protein structural predictions; phylogenetic cross-comparison with other *Plasmodium* species*
Differ significantly from probable human orthologues	Comparison with human sequences*
Apicoplast pathway enzymes	GO function and location assignments*, text search for 'chloroplast or plastid'*, phylogenetic comparison with plants
Expressed in blood-stage parasites	Expression profiling studies*; proteomics data*
Essential for parasite survival	Curated annotation from pharmacological/genetic studies; literature searches
Validated drug targets (in other systems)	Drug databases; literature databases
Availability of candidate inhibitors	Small-molecule databases; DOCKing algorithms
Vaccine candidates	
Known antigens (AMA1, MSP1)	Text*, motif* and BLAST* searches
Multigene families	Self-BLAST analysis*
Associated with the parasite on infected cell surface	Protein features: signal sequences*, transmembrane domains*, potential acylation sites or GPI anchors. Similarity to known membrane and surface proteins in other systems
Immunodominant	B- and/or T-cell epitope predictions*
Unlikely to be deleted	Non-telomeric*; conserved in multiple *P. falciparum* isolates
Likely to be under positive (immune) sele ction	DNA/protein features: repetitive/low complexitysequence*. High ratio of non-silent/silent polymorphisms (from phylogenetic cross-comparisons and population genetic studies)*

*Searches currently supported by PlasmoDB

For immunologists, computationally accessible queries enable identification of particular genes of interest as vaccine antigens (see Table 1). Additional gene-family members can be recognized on the basis of sequence similarity. Probable surface antigens can be identified from the presence of signal sequences, transmembrane domains, acylation signals or glycophosphatidylinositol (GPI) anchor motifs. Additional queries of immunological relevance might include the presence of predicted immunodominant epitopes, expression in life-cycle stage(s) of interest, conservation in multiple *P. falciparum* isolates, and evidence of immune selection based on highly repetitive elements, low-complexity sequence or polymorphisms identified in population genetic studies.

PlasmoDB can be used to build complex queries using boolean operators. For example, searching PlasmoDB release 3.3 for all genes predicted to contain a secretory signal sequence yields 1,952 hits. Because this search used curated annotations plus the predictions from any one of several distinct gene-finding algorithms, the results are several-fold redundant, yielding about 800 distinct genes, or more than 15% of the parasite genome. More than twice as many proteins (5,003) are predicted to contain transmembrane domains, but the intersection of these results yields only 1,083 hits (about 400 distinct proteins) exhibiting both features. Next, the database can be searched for all predicted genes known to be expressed from expressed sequence tag (EST) evidence, yielding 3,057 hits (searches based on microarray or proteomics evidence are also possible). The intersection between these secretory pathway and expression search results identifies a grand total of 190 candidates, probably corresponding to fewer than 100 distinct genes.

Two key points emerge from these queries. First, the power of a database devoted to mining genomics-scale data sets comes from its ability to form relational (integrated) queries, allowing researchers to frame their own questions. No encyclopaedic version of precomputed analyses and 'canned' queries will ever provide all possible answers in advance. For example, neither computational analysis nor manual curation would have been likely to identify enzymes associated with the apicoplast before this organelle was discovered and its targeting signals mapped.

Second, the goal of these queries is not to get the 'right' answer (a provably correct list of valid drug targets or vaccine antigens), but to reduce the range of options, filtering the overwhelming number of sequences in the genome down to a few genes amenable to experimental analysis — in short, to let computers do what computers do well, and to let people do what people do well. Integrating the results of such studies into the database completes the loop, with computational and laboratory analysis building on each other to accelerate the pace of biological research. ■

Jessica C. Kissinger, Brian P. Brunk, Jonathan Crabtree, Martin J. Fraunholz, Bindu Gajria, Arthur J. Milgram, David S. Pearson, Jonathan Schug, Amit Bahl, Sharon J. Diskin, Hagai Ginsburg, Gregory R. Grant, Dinesh Gupta, Philip Labo, Li Li, Matthew D. Mailman, Shannon K. McWeeney, Patricia Whetzel, Christian J. Stoeckert Jr and David S. Roos are associated with the Departments of Biology and Genetics, Center for Bioinformatics and Genomics Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018, USA. Address for correspondence: droos@sas.upenn.edu.