Published online 2 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D547–D552***
doi:10.1093/nar/gkm820

ProtozoaDB: dynamic visualization and exploration of protozoan genomes

Alberto M. R. Dávila[1,*], Pablo N. Mendes[9], Glauber Wagner[1,2,3], Diogo A. Tschoeke[1], Rafael R. C. Cuadrat[1], Felipe Liberman[1], Luciana Matos[4], Thiago Satake[11], Kary A. C. S. Ocaña[1], Omar Triana[1,10], Sérgio M. S. Cruz[4,5], Henrique C. L. Jucá[1], Juliano C. Cury[1], Fabricio N. Silva[4], Guilherme A. Geronimo[2], Margarita Ruiz[1], Eduardo Ruback[1], Floriano P. Silva Jr[1], Christian M. Probst[6], Edmundo C. Grisard[2], Marco A. Krieger[6], Samuel Goldenberg[6], Maria C. R. Cavalcanti[7], Milton O. Moraes[1], Maria L. M. Campos[8] and Marta Mattoso[4]

[1]Oswaldo Cruz Institute, FIOCRUZ, [2]CCB, Federal University of Santa Catarina (UFSC), [3]ACBS/UNOESC/SC, [4]COPPE, Federal University of Rio de Janeiro, [5]NCE, Federal University of Rio de Janeiro, [6]Instituto de Biologia Molecular do Paraná, FIOCRUZ, [7]Engineering Military Institute (IME), [8]DCC, Federal University of Rio de Janeiro, Brazil, [9]Knoesis Center, Wright State University, USA, [10]Antioquia University, Colombia and [11]Federal University of Paraná (UFPR), Brazil

Received August 15, 2007; Revised September 17, 2007; Accepted September 18, 2007

ABSTRACT

ProtozoaDB (http://www.biowebdb.org/protozoadb) is being developed to initially host both genomics and post-genomics data from *Plasmodium falciparum*, *Entamoeba histolytica*, *Trypanosoma brucei*, *T. cruzi* and *Leishmania major*, but will hopefully host other protozoan species as more genomes are sequenced. It is based on the Genomics Unified Schema and offers a modern Web-based interface for user-friendly data visualization and exploration. This database is not intended to duplicate other similar efforts such as GeneDB, PlasmoDB, TcruziDB or even TDRtargets, but to be complementary by providing further analyses with emphasis on distant similarities (HMM-based) and phylogeny-based annotations including orthology analysis. ProtozoaDB will be progressively linked to the above-mentioned databases, focusing in performing a multi-source dynamic combination of information through advanced interoperable Web tools such as Web services. Also, to provide Web services will allow third-party software to retrieve and use data from ProtozoaDB in automated pipelines (workflows) or other interoperable Web technologies, promoting better information reuse and integration. We also expect ProtozoaDB to catalyze the development of local and regional bioinformatics capabilities (research and training), and therefore promote/enhance scientific advancement in developing countries.

INTRODUCTION

Among the most important protozoan parasitic species causing diseases in humans, five had their genomes recently sequenced. Three of them are pathogenic members of the Order Kinetoplastida, Family Trypanosomatidae, that cause Sleeping sickness (caused by pathogenic subspecies of *Trypanosoma brucei*), Chagas disease (*T. cruzi*) or leishmaniasis (*Leishmania major*) representing the major human diseases caused by kinetoplastids (1–3). The fourth pathogenic species in question is *Plasmodium falciparum*, a member of the Phylum Apicomplexa and the causative agent of the most important human malaria (4). The last is Entamoeba histolytica, member of the Subphylum Sarcodina, the causative agent of amebiasis that infects approximately 50 million people over the globe, mainly on neotropical developing countries (5).

The availability of pathogenic protozoan genome sequences allows for comparative analyses to be performed in a systematic and straightforward way,

*To whom correspondence should be addressed. Tel: +55 21 3865 8229; Fax: +55 21 3865 8229; Email: davila@fiocruz.br

towards a better understanding of biological, genetic and evolutionary aspects. Also, such analyses allow the identification of genes related to pathogenesis and/or species-specific genes, useful for diagnostic markers as well as genes involved in crucial metabolic pathways that may lead to drug or vaccine development. Identification of homolog genes, particularly orthologs, became an important task to transfer annotation from experimentally or better characterized genes, allowing annotation and re-annotation of genes and genomes.

Considering the existing genomic data from important human protozoan parasites as well as ongoing sequencing efforts of several other relevant pathogenic species such as *P. vivax*, *Toxoplasma gondii*, *Cryptosporidium parvum*, *T. rangeli*, *L. braziliensis* and *L. chagasi* [GOLD (6)], ProtozoaDB is being developed to initially host genomics and post-genomics data from *P. falciparum*, *E. histolytica*, *T. brucei*, *T. cruzi* and *L. major* (5-protozoa), with added-value obtained from similarity- and phylogeny-based analyses performed as part of the system pipeline, but considering the adding/linking of more pathogenic species database in a near future.

Data Inventory and Analysis Tools

All 5-Protozoa nucleotide entries available at GenBank (release 160), RefSeq (release 24) and EST (release 160) divisions were downloaded from NCBI (Table 1) and loaded into ProtozoaDB using the Genomics Unified Schema (GUS: http://www.gusdb.org) system, version 3.5, for Postgres (http://www.postgres.org).

Redundancy of sequence proteins was treated in the following way: (i) the cd-hit (7) software was ran on all proteins of each five species, providing information on proteins with 100% identity, and (ii) redundancy mapped to already existing GUS tables, then users will be able to query and retrieve non-redundant datasets. External database IDs (from TcruziDB and PlasmoDB) mappings were incorporated into GUS, so that users can query and retrieve ProtozoaDB data using GenBank, TcruziDB or PlasmoDB IDs. Visitors can also click on any of the mapped external database IDs to be re-directed to PlasmoDB, TcruziDB, GeneDB or Superfamily databases, then look on extra annotations. Orthologous groups of the 5-protozoa are being identified using the OrthoMCL package (8). Similarity analyses were performed using all the proteins coded by the GenBank and RefSeq entries listed in Table 1, consisting of BlastP against the UniRef90 database (release 11) of the UniProt consortium (http://www.pir.uniprot.org/database/nref.shtml). Further similarity analyses using the Conserved Domain Database (9) and InterPro (10) are being progressively added to the database.

Table 1. Number of nucleotide sequences from each parasitic protozoan species retrieved from the GenBank, Refseq and EST databases[a]

Species	GenBank	RefSeq	EST
P. falciparum	10 992	5282	21 349
E. histolytica	859	11 591	20 404
L. major	474	1292	2191
T. brucei	1559	8812	5133
T. cruzi	3276	52 353	13 971

[a]Database search executed in July 2007.

ProtozoaDB is designed to offer a variety of query-based search tools that allow the user to perform an easy search of genes among the 5-protozoan genomes through (i) keyword, (ii) gene ID, (iii) product, (iv) protein motifs or (v) sequence type (coding sequences, mRNA, rRNA, tRNA, snRNA, snoRNA, transcript primary, precursor RNA and untranslated sequences) searches. Also, the Web user interface allows to view protozoan sequences separately or to compare them with each other. Individual chromosomes can be visualized using Gbrowse (http://www.gmod.org/wiki/index.php/GBrowse) that has been integrated to ProtozoaDB. Additionally, when a query is executed, information related to sequences' features and DNA, as well as translated and protein sequences may be obtained. When a gene (e.g. 'tubulin') is searched, the user can retrieve 'tubulin' entries originally available from GenBank or RefSeq, with their corresponding 'tags' or keyword annotation, and also associated papers published in PubMed. All users can contribute to the (re-)annotation of 'tubulin' adding more keywords to describe function and/or localization, and track further (re-) annotation of that particular 'tubulin' entry through the addition of sequence or gene RSS (Really Simple Syndication) to their preferred RSS reader (or even using iGoogle or Netvibes) (Figure 1).

ProtozoaDB also contains a collection of ESTs from different life cycle stages of the distinct species, allowing the user to perform queries on genes related to metacyclogenesis and/or to compare genes expressed in different parasite species or stages. Due to their particular biological and medical importance, a collection of ESTs from epimastigote, promastigote and amastigote forms of *T. cruzi* and *L. major* is available. Moreover, ESTs from gametocytes, schizont and asexual stages of *P. falciparum* are also available for comparison. An extensive documentation is being incorporated progressively, then users can benefit from the HowTo (user manual), FAQ (Frequently Asked Questions), GUS-Postgres-Wiki (details on the installation, configuration and debugging of GUS for Postgres) and ProtozoaDB roadmap (features to be added) in the left-side menu of the main page.

With respect to ProtozoaDB schema, a complete new subschema named 'Phylo' was designed and incorporated into GUS 3.5, extending it in order to store data from phylogenetic experiments, e.g. molecular phylogeny using distance and maximum likelihood, and profile-based phylogeny, as described by Theobald and Wuttke 2005 (11). The phylogeny-based analyses are being progressively incorporated into ProtozoaDB. Currently, users can perform queries in the phylogenetic trees constructed for the enzymes coded by mobile genetic elements and for the 18S rDNA (Figure 2). At this stage, only trees obtained by the Neighbor-Joining and distances methods are being stored in the system. The ultimate goal is to have all the 5-protozoan homologous genes identified, to support inferences and to store the phylogeny of all orthologous

and paralogous genes, as well as conserved domains to allow other studies such as horizontal gene transfer.

SYSTEM ARCHITECTURE AND IMPLEMENTATION

ProtozoaDB system is publicly available at http://www.biowebdb.org/protozoadb and offers modern Web interfaces, based on concepts inspired by what has been called the Web2.0 (http://en.wikipedia.org/wiki/Web_2) and REST-based Web services (http://en.wikipedia.org/wiki/Representational_State_Transfer). Those design principles can enhance e-scientists experiment perception, by improving the Web applications' look-and-feel, facilitating collaboration and sharing of resources, and, most of all, facilitating services interoperability. ProtozoaDB offers more flexibility than traditional (Perl) scripts while being more flexible and efficient than Web services technology. Yet, it remains compatible with Web services protocols as well as with scripts, turning the system extremely customizable.

ProtozoaDB system heavily relies on Web technology. We currently use the developed REST-based Web services as content providers for quick dynamic responses to user actions through the system interface, as well as to enable content and functionality sharing between ProtozoaDB and third-party websites. In general, ProtozoaDB interoperates in three distinct categories:

(i) Use of third-party Web services for dynamic integration of information: an example of the use of lightweight services is provided through a ProtozoaDB user interface, where we merge information obtained from ProtozoaDB and NCBI's Entrez Programming Utilities (e-Utils: http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html). In this specific example, we search Pubmed dynamically, providing links to literature whenever an article cites sequences stored in ProtozoaDB;

(ii) Provision of Web services for third-party reuse: a collection of services are offered to facilitate data retrieval in ProtozoaDB, intended to be used by automated third-party pipelines (workflows) or external user interfaces interested in interoperating with ProtozoaDB;

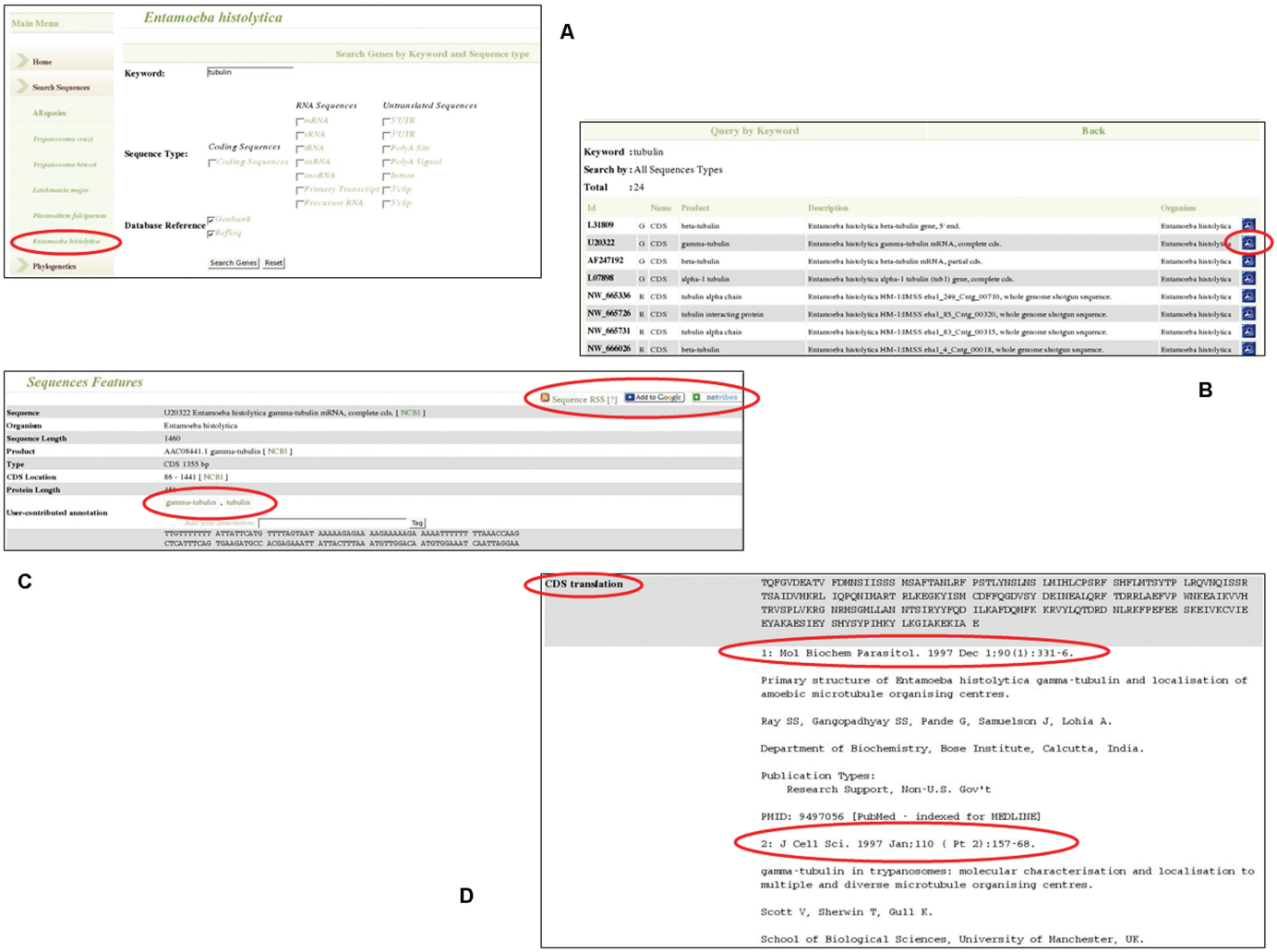

Figure 1. Features of the Search Sequences tool in ProtozoaDB. (**A**) searching *E. histolytica* subset, (**B**) choosing U20322 entry, (**C**) details of U20322 entry that can be tracked with RSS, Google or Netvibes and (**D**) CDS of U20322 entry and associated abstracts published in PubMed.

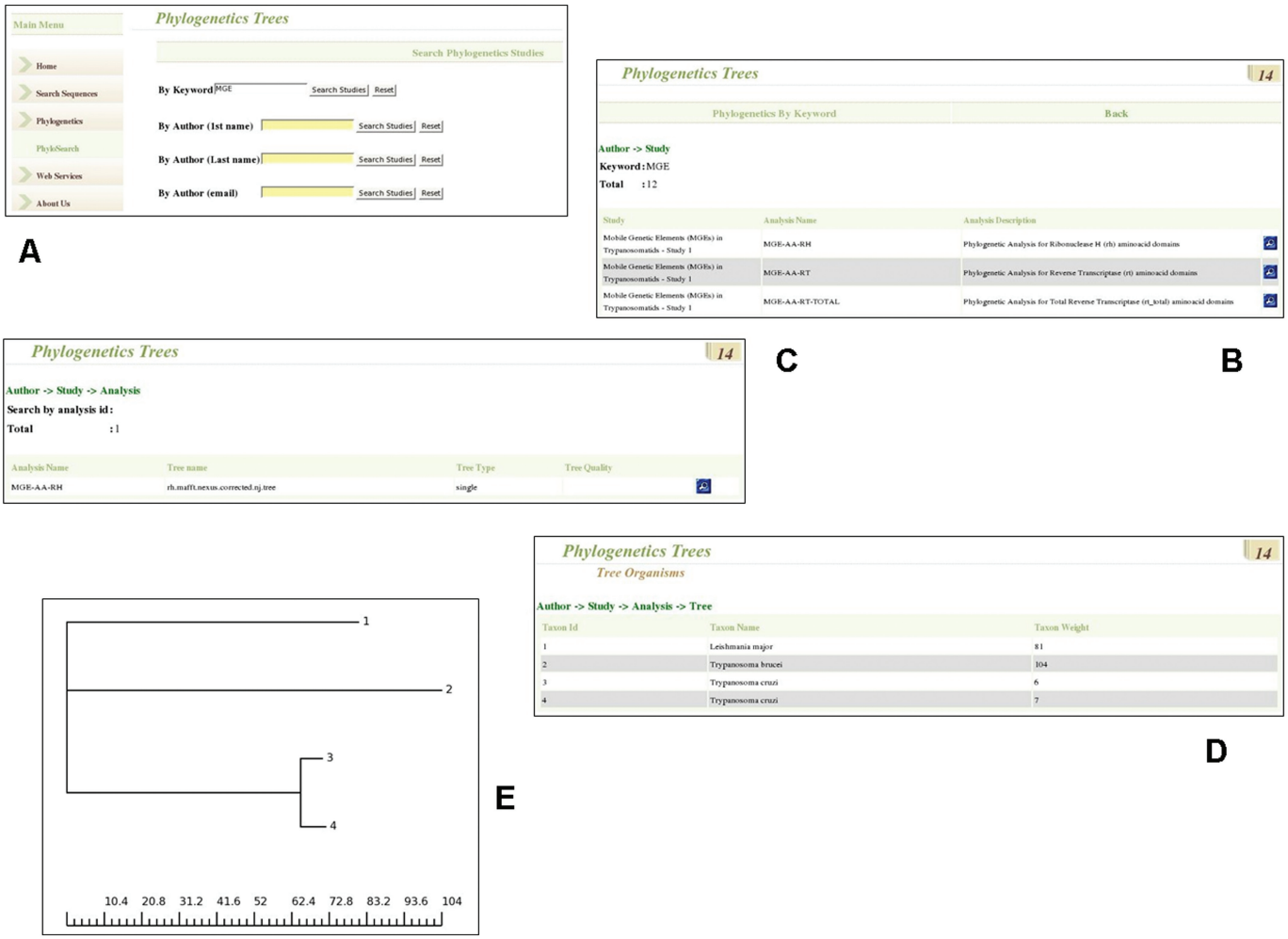

Figure 2. Features of the ProtozoaDB Phylo search tool. (**A)** searching Mobile Genetics Elements (MGE) phylogenetic trees, (**B)** MGE genes trees, (**C)** Ribonuclease H available data, (**D)** trypanosomatid taxons and (**E)** phylogenetic tree of Ribonuclease H.

(ii) Tools for direct community involvement: ProtozoaDB interface encourages users involvement by offering them the option of entering their amendments or full annotations to any sequence in the form of 'tags'—a concept successfully used in content-sharing websites such as Flickr (http://flickr.com/tour/) and Delicious (http://del.icio.us/), and more recently, NCBI Entrez with the introduction of GeneRIFs (Gene Reference Into Function: http://www.ncbi.nlm.nih.gov/projects/GeneRIF/). Additionally, an RSS feed option is available for user subscription, allowing him/her to receive notification of any updates on a given sequence of interest. RSS feeds are widely supported by current Web technology, such as e-mail clients (e.g. Mozilla Thunderbird), browsers (e.g. Mozilla Firefox), iGoogle (Google Customized Homepage) and Netvibes (http://www.netvibes.com).

The general architecture of ProtozoaDB system is basically composed by three main modules: Web Services, Query System and Database-Loading Plugins (GUS plugins) (Figure 3). The REST-based Web services constitute its main layer, through where all the services are provided and by which ProtozoaDB data can be accessed worldwide. A Web page containing a Javascript component provides a way for a dynamic user-friendly interaction with the ProtozoaDB services. Other application clients such as third-party pipelines can also directly connect to these services and reuse ProtozoaDB data. Alternatively, ProtozoaDB provides automatic database update notifications through RSS feeds to the users, reinforcing the community-based research aspect of the system. The GUS plugins layer performs a key role on the ProtozoaDB architecture. It aims to process genomic data originated from external data sources (e.g. GenBank, OBO, etc.) and to feed such data into the ProtozoaDB (GUS-based extended schema). GUS plugins are open-source Perl programs and are bundled with the GUS distribution kit. Being open source, GUS facilitates distinct projects to write and share personalized plugins, enhancing research projects/teams collaboration through its 'standard' common database schema. The ProtozoaDB architecture allows data storage in a central repository that locally integrates processed core genomic data, but a loosely coupled integration with external sources is envisioned and preliminarily supported. We believe that

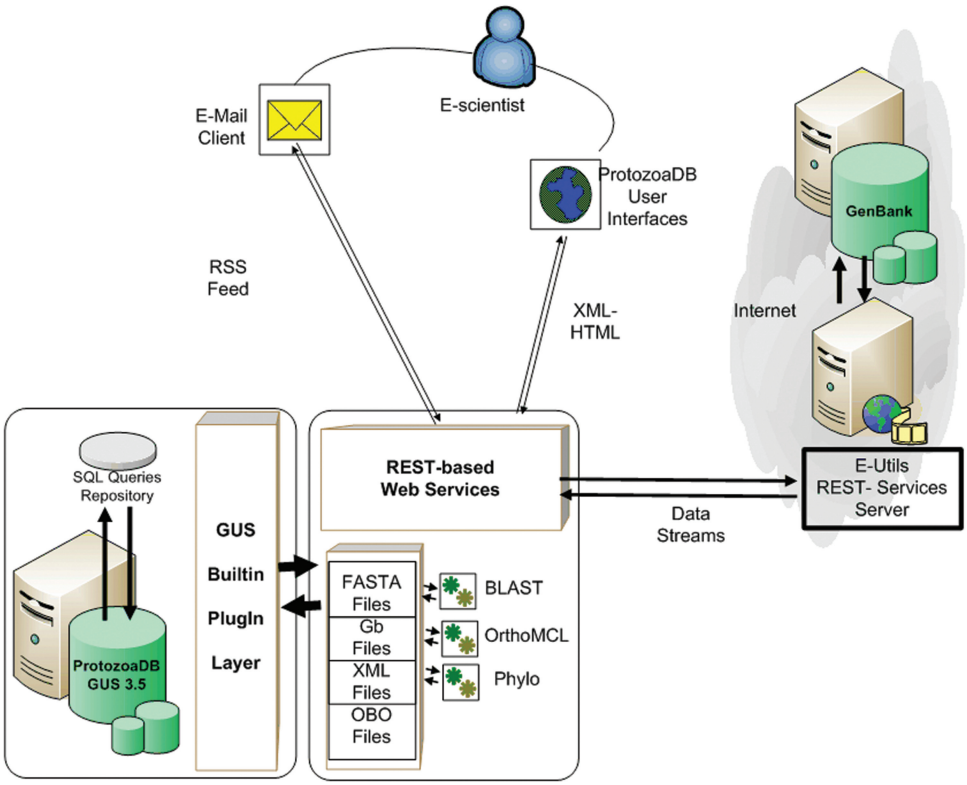

Figure 3. General overview of the ProtozoaDB Architecture.

in the near future, other initiatives will also adopt content sharing via Web services, allowing a more flexible and dynamic integration with ProtozoaDB and thus offering a user-friendly and powerful database.

In addition to Web services access, the ProtozoaDB database is available through a user interface, where a series of pre-defined database queries is offered to the users in order to ease the analysis task. Finally, the eUtils (Entrez Programming Utilities) module, provided by NCBI, which does not currently integrate the ProtozoaDB architecture, was included in Figure 3 to illustrate the system ability to provide services composition with third-party database sites.

FUTURE DEVELOPMENTS

Our plans include: (i) design and incorporation of another new subschema into GUS 3.5 for protein structure experiments, including 'druggability' inferences; (ii) link/interactions with TcruziDB (12), ApiDB (13), GeneDB (http://www.genedb.org) and TDRtargets (http://www.tdrtargets.org) databases; (iii) inclusion of sequences from more protozoan species whose genome/transcriptome/proteome data generation are currently under way; (iv) the addition of post-genomics data from microarrays, proteomics and 3D structures experiments; (v) at least a yearly release cycle for ProtozoaDB and (vi) the inclusion of a non-restrictive Creative Commons license.

In step with the emergence of the Semantic Web, an envisioned third generation of the Web, our database was partially mapped to a set of ontologies. Providing accessibility to our repository through Semantic Web-based technologies, we offer Web access to formal relationships that can be explored both by humans and machines, towards a Relationship Web (14).

ACKNOWLEDGEMENTS

To CNPq (Brazilian Research Council), PAPES-IV/FIOCRUZ, MCT/MS-SCTIE-DECIT, Fogarty International Center (Grant number: 5D43TW007012-03), FAPERJ and TWAS for financial support. E.C.G. is currently a CNPq Postdoctoral Fellow at BMRC/UEA, UK. O.T. is a CNPq/TWAS postdoctoral fellow at the Institute Oswaldo Cruz; D.A.T., M.R. and H.C.L.J. are TecTec (Fiocruz/Faperj) fellows at the Instituto Oswaldo Cruz. To Guilherme de Oliveira and his group for sharing their unpublished results on their local version of GUS. To the GUS, BioPerl and Open Source communities for being so helpful. To Mark Heiges and Jessica Kissinger for PlasmoDB and TcruziDB ID mappings. Funding to pay the Open Access publication charges for this article was provided by Oswaldo Cruz Institute, FIOCRUZ.

Conflict of interest statement. None declared.

REFERENCES

1. Berriman,M., Ghedin,E., Hertz-Fowler,C., Blandin,G., Renauld,H., Bartholomeu,D.C., Lennard,N.J., Caler,E., Hamlin,N.E. *et al.* (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science*, **309**, 416–422.
2. El-Sayed,N.M., Myler,P.J., Bartholomeu,D.C., Nilsson,D., Aggarwal,G., Tran,A.N., Ghedin,E., Worthey,E.A.,

Delcher,A.L. *et al*. (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science*, **309**, 409–415.
3. Ivens,A.C., Peacock,C.S., Worthey,E.A., Murphy,L., Aggarwal,G., Berriman,M., Sisk,E., Rajandream,M.A., Adlem,E. R. *et al*. (2005) The genome of the kinetoplastid parasite, *Leishmania major*. *Science*, **309**, 436–442.
4. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E. *et al*. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498–511.
5. Loftus,B., Anderson,I., Davies,R., Alsmark,U.C., Samuelson,J., Amedeo,P., Roncaglia,P., Berriman,M., Hirt,R.P. *et al*. (2005) The genome of the protist parasite *Entamoeba histolytica*. *Nature*, **433**, 865–868.
6. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res*., **34**, D332–D334.
7. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
8. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*., **13**, 2178–2189.
9. Marchler-Bauer,A., Anderson,J.B., Derbyshire,M.K., DeWeese-Scott,C., Gonzales,N.R., Gwadz,M., Hao,L., He,S., Hurwitz,D.I. *et al*. (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res*., **35**, D237–D240.
10. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L. *et al*. (2007) New developments in the InterPro database. *Nucleic Acids Res*., **35**, D224–D228.
11. Theobald,D.L. and Wuttke,D.S. (2005) Divergent evolution within protein superfolds inferred from profile-based phylogenetics. *J. Mol. Biol*., **354**, 722–737.
12. Aguero,F., Zheng,W., Weatherly,D.B., Mendes,P. and Kissinger,J.C. (2006) TcruziDB: an integrated, post-genomics community resource for *Trypanosoma cruzi*. *Nucleic Acids Res*., **34**, D428–D431.
13. Aurrecoechea,C., Heiges,M., Wang,H., Wang,Z., Fischer,S., Rhodes,P., Miller,J., Kraemer,E., Stoeckert,C.J. Jr *et al*. (2007) ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res*., **35**, D427–D430.
14. Sheth,A. and C. Ramakrishnan (2007) Relationship Web: Blazing Semantic Trails between Web Resources. *IEEE Internet Comput*., **11**, 77–81.