TcruziDB, an Integrated Database, and the WWW Information Server for the *Trypanosoma cruzi* Genome Project

Wim Degrave/+, Antonio B de Miranda, Alex Amorim, Adeílton Brandão, Martin Aslett*, Mark Vandeyar**

Departamento de Bioquímica e Biologia Molecular, Instituto Oswaldo Cruz, Av. Brasil 4365, 21045-900 Rio de Janeiro, RJ, Brasil *European Bioinformatics Institute (EBI), Hinxston, UK **Av. Couronnement 27, 1200 Brucels, Belgium

Data analysis, presentation and distribution is of utmost importance to a genome project. A public domain software, ACeDB, has been chosen as the common basis for parasite genome databases, and a first release of TcruziDB, the Trypanosoma cruzi *genome database, is available by ftp from ftp://iris.dbbm.fiocruz.br/pub/genomedb/TcruziDB as well as versions of the software for different operating systems (ftp://iris.dbbm.fiocruz.br/pub/unixsoft/). Moreover, data originated from the project are available from the WWW server at http://www.dbbm.fiocruz.br. It contains biological and parasitological data on CL Brener, its karyotype, all available* T. cruzi *sequences from Genbank, data on the EST-sequencing project and on available libraries, a* T. cruzi *codon table and a listing of activities and participating groups in the genome project, as well as meeting reports.* T. cruzi *discussion lists (tcruzi-l@iris.dbbm.fiocruz.br and tcgenics@iris.dbbm.fiocruz.br) are being maintained for communication and to promote collaboration in the genome project.*

Key words: TcruziDB - genome database - *Trypanosoma cruzi* - clone CL Brener - Parasite Genome Projects - anonymous ftp - listserver

Parasite genome projects, some of which started in 1994 after the WHO/TDR sponsored Parasite Genome Network Planning Meeting at Fiocruz (Rio de Janeiro, Brazil, 14-15 April 1994), can be considered as highly successful. Scientists from developing and developed countries have planned and initiated the projects and several “consortiums” for the mapping and sequencing of these medium-sized genomes were established, often based on already ongoing North-South and South-South collaborations. Thus, the genomes of *Plasmodium falciparum, Schistosoma mansoni*, *Trypanosoma cruzi*, *Leishmania major*, *Trypanosoma brucei, Brugia malayi*, amongst others, are now under study (see http://www.ebi.ac.uk/parasites/parasite-genome.html). The main objectives for these projects are to increase drastically the knowledge on the (molecular) biology of these parasites, to identify new genes with key cellular functions, which could be eligible as target for new drugs or new antigens for use in diagnostics or vaccine development, and to promote technology transfer and cooperation between the different participating laboratories.

The *T. cruzi* genome initiative has been working along these lines, and recent reviews on scientific progress have been published (Degrave et al. 1997, Zingales et al. 1997). More than 20 collaborating centers are contributing to the initiative, of which the majority are located in Latin-America, where Chagas disease is endemic. During the first two years of the project, emphasis has been laid on the structuring of the network, acquisition of adequate basic infrastructure for genome research and expertise in new techniques, construction of genomic and cDNA libraries and further characterization of the parasite, including its karyotype. The project has now entered into a large scale sequencing phase, both on ESTs and selected chromosomes, and we can expect the generation of large quantities of new data during the next few years.

However, one of the critical activities in a genome project deals with bioinformatics. It is widely recognized that this field, usually somewhat neglected, will become a major area of study in the

Financial support: UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases; CYTED-Subprogram of Biotechnology (Spain), CNPq, Centro Brasileiro Argentino de Biotecnologia (CBAB/CABBIO).

+Corresponding author. Fax: +55-21-590.3495. E-mail: wdegrave@gene.dbbm.fiocruz.br

Received 20 August 1997
Accepted 10 September 1997

next decade. Indeed, information, such as nucleotide sequences and maps, are quite useless if it is not analyzed, processed and presented to the greater scientific community.

RESULTS AND DISCUSSION

TcruziDB, an integrated database for the T. cruzi genome project - At the WHO/NCBI Parasite Genome Computing Workshop, held at Woods Hole Marine Biological Laboratory (Sept. 14-16, 1995), it was decided that the ACeDB software, an object oriented database manager specifically developed for large-scale molecular biology, such as genome projects, would be adopted for the parasite genome projects. The ACeDB software was written initially for *the Caenorhabditis elegans* project (Durbin & Thierry-Mieg 1991), but has since then been adapted for a great number of other projects (for an overview, see the ACeDB-faq at http://probe.nalusda.gov/acedbfaq.html), including several plant and parts of the human genome project. ACeDB is an object-oriented database with tools for displaying and analyzing most of the types of objects used in molecular biology and genome projects, such as genes, antigens, genetic maps, sequences, clones, contigs, filter grids, literature data and authors, collaborating colleagues etc. The database can be adapted for the specific needs and information from a certain genome project, and data can be cross referenced where needed according to the "model". User guides can be obtained from http://probe.nalusda.gov:8000/acedocs/. A major drawback of the ACeDB software is that it runs on either UNIX or Linux systems or on Macintosh computers, under X windows, hampering its widespread use in developing countries. The recent release of Winace, a version running on personal computers with the Win95 or WinNT operating system, should be considered as a major breakthrough.

The first version of the TcruziDB v 1.1 genome database in ACeDB format (version 4.3i) was created and currently contains the following data (Fig. 1):

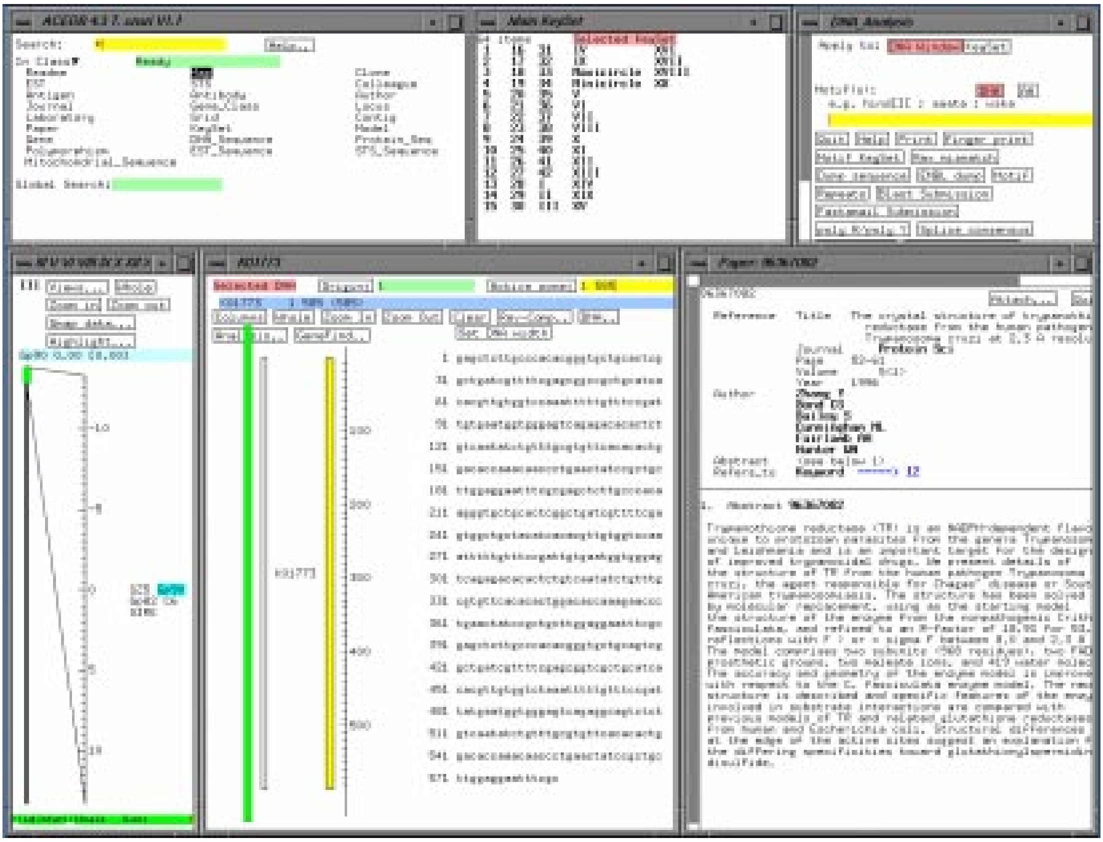


Fig. 1: typical screen of TcruziDB. In the upper left: the main panel with clickable objects. Upper middle: main keyset screen, showing elements of the map object (chromosomal bands in this case). Upper right: the DNA analysis command screen. Lower left: graphical representation of a chromosome. Lower middle: graphical representation of a chromosome fragment and the nucleotide sequence of a mapped gene. Lower right: typical screen of a bibliographic reference.

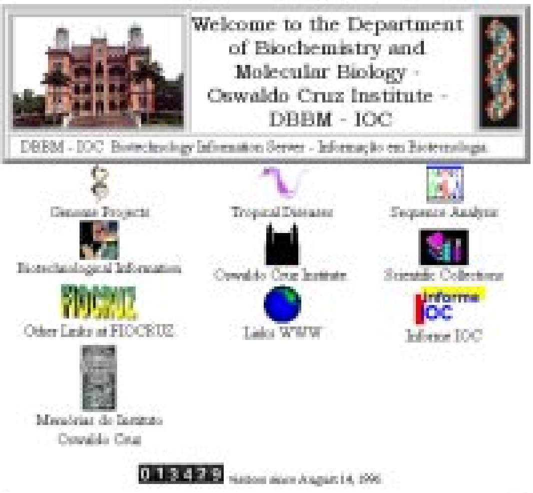
- *T. cruzi* CL Brener chromosomes (numbered I-XX and 1-42, to reflect the karyotype analyses), as well as minicircle/maxicircle molecules
- Mapped loci (50 entries)
- *T. cruzi* sequences from GB and EMBL (279 entries)
- *T. cruzi* medline references since 1966 (3573 entries)
- Addresses and data on the collaborators in the *T. cruzi* genome initiative
- *T. cruzi* protein data from Swissprot.

Ace files for data entry were created through awk and modification of existing perl scripts. The “model” used for this release is similar to the one used for LeishDB, with some modifications. TcruziDB, as well as the databases from the other parasite genome projects, is available via ftp at ftp://iris.dbbm.fiocruz.br/pub/genomedb/TcruziDB. ACeDB software for a variety of operating systems (UNIX, Linux and also Winace) are available at the same site in the directory /pub/unixsoft/.

A second version of TcruziDB v2.1 will be available by the time of this publication and will include updates on the sequences (650 entries), EST sequences (1100 entries or more), medline references, protein data from Swissprot and TREMBL and on collaborators, and will also include new data on filter grids for cosmid and BAC libraries, and fasta alignments. Version 2.1 has modifications in the “model”, in order to include pictures and particular data from the *T. cruzi* project.

The DBBM/IOC biotechnology and genome information server - The WWW server of the Department for Biochemistry and Molecular Biology/Oswaldo Cruz Institute (http://www.dbbm.fiocruz.br; Fig. 2) offers information and links on genome projects, on some tropical diseases and on general biotechnology, as well as on nucleotide and protein sequence analysis, and provides some general services.

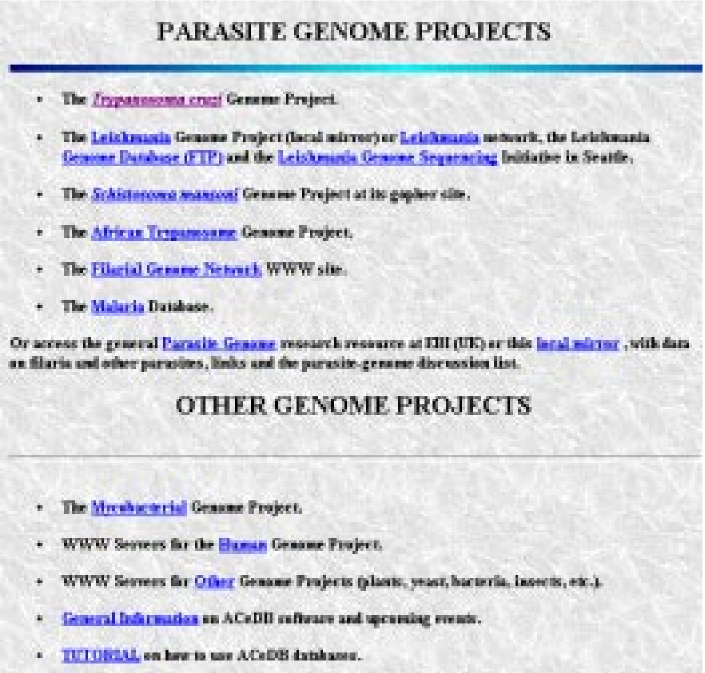
Besides offering links to all major (parasite) genome projects (Fig. 3, Table), the server is the central site for the *T. cruzi* genome project (Fig. 4), be-



Welcome to the Department of Biochemistry and Molecular Biology - Oswaldo Cruz Institute - DBBM - IOC

DBBM - IOC Biotechnology Information Server - Informação em Biotecnologia

Fig. 2: main page of the DBBM/IOC WWW server.



PARASITE GENOME PROJECTS

- The *Trypanosoma cruzi* Genome Project.
- The Leishmania Genome Project (local mirror) or Leishmania network, the Leishmania Genome Database (FTP) and the Leishmania Genome Sequencing Initiative in Seattle.
- The *Schistosoma mansoni* Genome Project at its gopher site.
- The African Trypanosome Genome Project.
- The Filarial Genome Network WWW site.
- The Malaria Database.

Or access the general Parasite Genome research resource at EBI (UK) or this local mirror , with data on filaria and other parasites, links and the parasite-genome discussion list.

OTHER GENOME PROJECTS

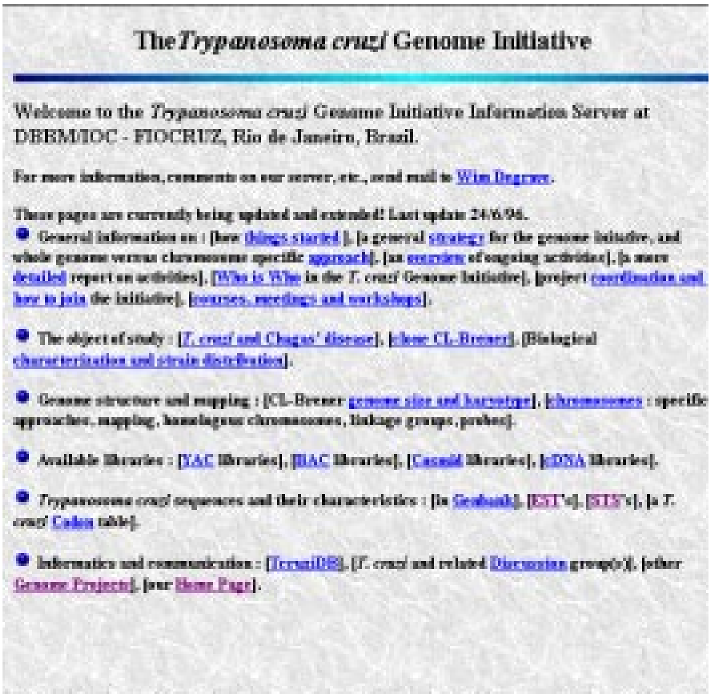
- The Mycobacterial Genome Project.
- WWW Servers for the Human Genome Project.
- WWW Servers for Other Genome Projects (plants, yeast, bacteria, insects, etc.).
- General Information on ACeDB software and upcoming events.
- TUTORIAL on how to use ACeDB databases.

Fig. 3: WWW page with links to genome project web pages.

TABLE

Internet access to parasite genome project information servers

Parasite Genome Central Resources	
WWW site	http://www.ebi.ac.uk/parasites/parasite-genome.html
email network	parasite-genome@mailbase.ac.uk
ftp site	ftp://ftp.ebi.ac.uk/pub/databases/parasites
Trypanosoma cruzi Genome Project	
WWW sites	http://www.dbbm.fiocruz.br/genome/tcruzi/tcruzi.html http://www.dbbm.fiocruz.br/tropical/chagas/trypan.html
email network	tcruzi-l@iris.dbbm.fiocruz.br (general) tcgenics@iris.dbbm.fiocruz.br (closed list)
ftp site	ftp://iris.dbbm.fiocruz.br/pub/genomedb/TcruziDB
Malaria (Plasmodium) Genome Project	
Network WWW sites	http://www.wehi.edu.au/biology/malaria/who.html http://www.wehi.edu.au/biology/malaria/wellcome.html
email network	malaria@wehi.edu.au
Trypanosoma brucei Genome Project	
WWW site	http://parsun1.path.cam.ac.uk/newtryp/toppage.htm
ftp site	ftp://ftp.ebi.ac.uk/pub/databases/parasites/brucei/
Leishmania Genome Project	
Genome Project WWW site (UK)	http://www.ebi.ac.uk/parasites/leish.html
Genome Project WWW site (US)	http://chimera.biotech.washington.edu/lgnsea.htm
email network	LeishL@bdt.org.br
ftp site	ftp://ftp.ebi.ac.uk/pub/databases/parasites/Leish/
Filarial Genome Project	
Genome Project WWW site	http://helios.bto.ed.ac.uk/mbx/fgn/filgen.html
email network	filarial-genome@mailbase.ac.uk
ftp site	ftp://ftp.ebi.ac.uk/pub/databases/parasites/Brugia/
Schistosoma Genome Project	
WWW site	http://www.nhm.ac.uk/schisto/
email network	bionet.organisms.schistosoma schistosoma@dl.ac.uk
ftp site	ftp://ftp.ebi.ac.uk/pub/databases/parasites/Schisto/



The *Trypanosoma cruzi* Genome Initiative

Welcome to the *Trypanosoma cruzi* Genome Initiative Information Server at DBBM/IOC - FIOCRUZ, Rio de Janeiro, Brazil.

For more information, comments on our server, etc., send mail to Wim Degrave.

These pages are currently being updated and extended! Last update 24/6/96.

- General information on : [how things started], [a general strategy for the genome initiative, and whole genome versus chromosome specific approach], [an overview of ongoing activities], [a more detailed report on activities], [Who is Who in the *T. cruzi* Genome Initiative], [project coordination and how to join the initiative], [courses, meetings and workshops].
- The object of study : [*T. cruzi* and Chagas' disease], [clone CL-Brener], [Biological characterization and strain distribution].
- Genome structure and mapping : [CL-Brener genome size and karyotype], [chromosomes : specific approaches, mapping, homologous chromosomes, linkage groups, probes].
- Available libraries : [YAC libraries], [BAC libraries], [Cosmid libraries], [cDNA libraries].
- *Trypanosoma cruzi* sequences and their characteristics : [in Genbank], [EST's], [STS's], [a *T. cruzi* Codon table].
- Informatics and communication : [TcruziDB], [*T. cruzi* and related Discussion group(s)], [other Genome Projects], [our Home Page].

Fig. 4: main page with information on the *Trypanosoma cruzi* genome project.

sides mirroring the sites for the project on *Leishmania* and for the parasite-genome pages. The pages on the *T. cruzi* genome project present general information, as well as details on available libraries, sequences, and EST data from our laboratory, before deposit to dbEST. Data for the web pages have been gathered through personal contact with the participants of the genome project.

Discussion lists, related to T. cruzi - A *T. cruzi* discussion list, automated through the majordomo software (subscription by sending a message "subscribe tcruzi-l"to majordomo@iris. dbbm.fiocruz.br; messages to tcruzi-l@iris.dbbm. fiocruz.br) is unmoderated and a monthly archive is made. Up to now, about 150 researchers from more than 20 countries are subscribed. The (closed) list tcgenics@iris.dbbm. fiocruz.br has been set up in order to improve the cooperation between the participating groups in the *T. cruzi* genome initiative, and deals specifically with technical communication on the genome project.

REFERENCES

Degrave W, Levin MJ, da Silveira JF 1997. Parasite genome projects and the *Trypanosoma cruzi* genome initiative. *Mem Inst Oswaldo Cruz* (this volume).

Durbin R, Thierry-Mieg J 1991. A *C. elegans* database. Documentation, code and data available from anonymous ftp servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.

Zingales B, Rondinelli E, Degrave W, da Silveira JF, Levin M, Le Paslier D, Modabber F, Dobrokhotov B, Swindle J, Kelly JM, Aslund L, Hoheisel JD, Ruiz AM, Cazzulo JJ, Pettersson U, Frasch AC 1997. The *Trypanosoma cruzi* genome initiative. *Parasitol Today 13*: 16-22.