ProtozoaDB: comparative genomics approaches for the study of protozoan species

Alberto M. R. Dávila
Email: davila@fiocruz.br

Laboratório de Biologia Computacional e Sistemas
Instituto Oswaldo Cruz - FIOCRUZ, Rio de Janeiro, Brasil.
Pathogenic protozoan species are commonly studied because they are a major cause of mortality and morbidity in humans and domestic animals throughout the world.

We have developed an automatic methodology to reconstruct the phylogenetic species tree in Protozoa integrating different phylogenetic algorithms and programs.

We demonstrate the utility of a supermatrix approach to construct phylogenomic trees using 31 universal representative orthologs.

These universal orthologs are involved in metabolic pathways, thus might potentially provide some extra information for the understanding of the phylogenetic relationships and evolutionary processes of Protozoa.
Illness caused by parasitic Protozoa are a major cause of disease worldwide, but because they are concentrated in low socio-economic parts of the world, they receive relatively little attention from the pharmaceutical industry and major scientific funding agencies.

Of the ten diseases targeted as research priorities by the World Health Organization’s Special Program for Research and Training in Tropical Diseases (http://www.who.int/tdr) four are caused by protozoan parasites (malaria, leishmaniasis, Chagas disease, African tripanosomiasis).

These diseases and other less dangerous ones (e.g. amoebiasis and trichomoniasis) are having an alarming increase of refractoriness cases to the main treatment. Treatment failure has potentially a multifactorial origin with one of the major concerns being drug resistance.

Phylogenomics is being intensively used in the “Post Genomic Era”. Many different phylogenetic trees have been published on the basis of different models of sequence evolution and utilization of different parameter settings and algorithms.
Although, the rRNA and other single genes have been extremely valuable for phylogenetic studies but they have their limitations.

Chaudhary (2005) and Keeling (2005) showed results on available sequencing data.

Phylogenetic analysis defined five supergroups of eukaryotes: (i) the plant and red/green algal lineage; (ii) a clade comprised of animals, fungi, slime molds and amebozoans; and three groups that are entirely protozoan: (iii) chromalveolates, (iv) excavates and (v) rhizaria.
Universal Orthologs:


Toward Automatic Reconstruction of a Highly Resolved Tree of Life

Francesca D. Ciccarelli,1,2,3* Tobias Doerks,1* Christian von Mering,1 Christopher J. Creevey,1 Berend Snel,4 Peer Bork1,5†

We have developed an automatable procedure for reconstructing the tree of life with branch lengths comparable across all three domains. The tree has its basis in a concatenation of 31 orthologs occurring in 191 species with sequenced genomes. It revealed interdomain discrepancies in taxonomic classification. Systematic detection and subsequent exclusion of products of horizontal gene transfer increased phylogenetic resolution, allowing us to confirm accepted relationships and resolve disputed and preliminary classifications. For example, we place the phylum Acidobacteria as a sister group of δ-Proteobacteria, support a Gram-positive origin of Bacteria, and suggest a thermophilic last universal common ancestor.

Reconstructing the phylogenetic relationships among all living organisms is one of the fundamental challenges in biology. Numerous attempts to derive a tree of life using various methods have been published [for a review, see (1)], and its principal existence has been questioned recently (2, 3). Moreover, even under the assumption of a tree of life, numerous groupings and taxonomic entities still remain heavily debated, and the advent of molecular and genomic data has increased the variety of classifications rather than reducing the problem (1).

Theoretical and practical limits to reconstructing a tree of life have been put forward, such as the insufficient amount of discriminating characters available, even in information-rich genomic data sets (4), and the computing resources required to cope with large numbers of species (1). Furthermore, there are factors that hamper accurate reconstruction of phylogenetic trees regardless of the methods used, such as sampling biases of species included (5) and dilution of phylogenetic signal by horizontal gene transfer (HGT) (6), the

---

1European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany. 2Department of Experimental Oncology, European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy. 3Institute of Molecular Oncology Foundation, Fondazione Italiana per la Ricerca sul Cancro, Via Adamello 16, 20139 Milan, Italy. 4Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Nijmegen Medical Center, Toernooiveld 1, 6525 ED, Nijmegen, Netherlands. 5Max Delbrück Centrum, 13092 Berlin-Buch, Germany.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: peer.bork@embl.de
Selection and preparation of marker gene families

- We based our methodology on Ciccarelli et al. (2006) study for the selection and construction of the species tree.

- Thirty-one universal orthologous (UO) genes were downloaded in fasta format from GenBank and then aligned using MAFFT (version 5.861)

Selection and preparation of marker gene families for protozoan species trees

- **Hidden Markov Models** (HMM) profiles were constructed for the 31 UO (local database) set and queried against all **Protozoa sequences** available in Refseq (RefSeq-release35.catalog-01/11/2009) and Genbank (NCBI-Flat File Release 172.0-06/15/2009).

- HMM profiles were created (**hmmbuild**) and calibrated (**hmmcalibrate**) using HMMER version 2.3.2 and searches done using **hmmpfam** with e-value “1e-5” as cut-off.

- Those best hmmpfam hits of each of the 31 UO against protozoan genomes were used to construct individual multiple alignments (Mafft v5.861).
Methodology 1 (M1):

- Those individual alignments were concatenated, then resulting in a global supermatrix of 21,260 positions in a total of 74 species.

- This supermatrix was used to generate a tree with PHYML using 100 bootstrap replicates and the JTT matrix.

Methodology 2 (M2):

- Those individual alignments were trimmed using TrimAl v1.2 aiming to eliminate the most variable regions of each of them. Those remaining conserved blocks were then concatenated, resulting in a global supermatrix of 12,807 positions in a total of 74 species.

- This supermatrix was used to generate a tree with PHYML using 100 bootstrap replicates and the JTT matrix.
Figure 2. Algorithm for Supermatrix versus Supertree construction.
We reconstructed the protozoan species tree using several automated bioinformatics tools: downloaded 31 universal UO from GenBank, built HMM profiles with them, searched those profiles in protozoan genomes (43 species) and related species (31 species) showing different level of completeness/coverage.

The species tree was formed by 3 major clades that were related to 3 groups of data:

(i) 26 species composed of at least 80% of UO (25/31) called \(G_1\),
(ii) 12 species composed of 50-79% (15-24/31) of UO called \(G_2\) and
(iii) 36 species composed of less than 50% (1-14/31) of UO called \(G_3\).
Results

Figure 1. Phylogenomic supermatrix radiation tree of protozoan using masked sequences with TrimAl. Was used a supermatrix of 12,807 positions in the total 74 protozoan species. Maximum likelihood tree was constructed with Phym, JTT as evolutionary model and bootstrap 100. (C1 is red, C2 blue and C3 black).
Results

- C1 was composed by only protozoan species, C2 was composed by rodophyta, cryptophyta, glaucocystophyceae, stramenopiles species, and C3 was composed by some species of C1 and C2.

- C1 presented monophyly of Protozoa, C2 presented monophyly of species related to Protozoa, and C3 polytomy of some Protozoa and related species. The more UO included in the supermatrix the better was the tree robustness inside the 3 clades.

- Figure 2 shows the comparison of trees using no masked (M1) and TrimAl masked (M2) sequences. Both methodologies showed very similar topologies, however the tree utilizing TrimAl masked sequences showed higher bootstrap values. Figure 1 is the same tree using TrimAl masked sequences but presented as radiation tree.
Supermatrix tree built with total sequences

Supermatrix tree built with conserved blocks

Results

Figure 2. Phylogenomic supermatrix trees of protozoan using total sequences and conserved blocks (TrimAl) sequences. Were used supermatrixes of 12,807 and 21,260 positions respectively in 74 species. Maximum likelihood tree was constructed with Phyml, JTT as evolutionary model and bootstrap 100. (C1 is red, C2 blue and C3 black).
Most variable regions trimmed (>80 UO): Clade 1 (zoom)
We have presented a phylogenomic-based overview for Protozoa. Relationships between protozoan groups are in agreement with previous studies, showing monophyly for protozoan.

On the other hand, phylogenetic information inferred from C3 is questionable due to incomplete genomes, suggesting that using less than 15 universal UO for phylogenomic reconstruction is not reliable.

The inclusion of more data/genes is necessary to obtain a robust tree.

Our phylogenomic-based methodology using a supermatrix approach proved to be reliable with protozoan genome data, suggesting that (a) more universal UO used the better, and (b) using the entire UO sequence or just a conserved block of it produce similar reliable results.

Finally, we need to investigate further if this methodology could be extrapolated or reproduced to other taxonomic groups.
ProtozoaDB: dynamic visualization and exploration of protozoan genomes

Alberto M. R. Dávila¹*, Pablo N. Mendes⁹, Glauber Wagner¹,²,³,
Diogo A. Tschoeke¹, Rafael R. C. Cuadrat¹, Felipe Liberman¹, Luciana Matos⁴,
Thiago Satake¹¹, Kary A. C. S. Ocaña¹, Omar Triana¹,¹⁰, Sérgio M. S. Cruz⁴,⁵,
Henrique C. L. Jucá¹, Juliano C. Cury¹, Fabricio N. Silva⁴,
Guilherme A. Geronimo², Margarita Ruiz¹, Eduardo Ruback¹,
Floriano P. Silva Jr¹, Christian M. Probst⁶, Edmundo C. Grisard²,
Marco A. Krieger⁶, Samuel Goldenberg⁶, Maria C. R. Cavalcanti⁷, Milton O. Moraes¹,
Maria L. M. Campos⁸ and Marta Mattoso⁴

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

Received August 15, 2007; Revised September 17, 2007; Accepted September 18, 2007
Approaches involving the integrative use of heterogeneous databases, analyses tools, distributed annotation systems as well as sensitive similarity detection algorithms have been catalyzed by a variety of sequenced genomes. In this purpose, the BiowebDB Consortium, partially funded by CNPq, present the Protozoa Database, an integrated, user-friendly and flexible platform that contains the several protozoa genomes: Trypanosoma cruzi, Trypanosoma brucei, Trypanosoma rangeli, Leishmania major, Plasmodium falciparum and Entamoeba histolytica.

Latest Version: 1.0 beta (August 2010)
## Query Results

<table>
<thead>
<tr>
<th>Accession Number</th>
<th>Name</th>
<th>Organism</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAB86601.1</td>
<td>cyclophilin</td>
<td>Entamoeba histolytica</td>
<td>details</td>
</tr>
<tr>
<td>AAC24632.1</td>
<td>cyclophilin</td>
<td>Leishmania major</td>
<td>details</td>
</tr>
<tr>
<td>AAC41390.1</td>
<td>cyclophilin</td>
<td>Plasmodium falciparum</td>
<td>details</td>
</tr>
<tr>
<td>AAC46975.1</td>
<td>cyclophilin</td>
<td>Plasmodium falciparum</td>
<td>details</td>
</tr>
<tr>
<td>AAD55769.1</td>
<td>cyclophilin</td>
<td>Plasmodium falciparum</td>
<td>details</td>
</tr>
<tr>
<td>AAF85986.1</td>
<td>cyclophilin A</td>
<td>Trypanosoma cruzi</td>
<td>details</td>
</tr>
</tbody>
</table>

## Details

- **ProtozoaDB**
  - Orthologs by OrthoSearch [go to paper]
  - Conserved Domains (Top 10 rpsBlast)
  - Human Proteome Similarity (Top 10 Blastp)
  - Orthologs by OrthoMCL [go to paper]
  - Protozoa EST Similarity (Top 10 Blastx)
## Query Results

### Accession Number | Name | Organism | Details
--- | --- | --- | ---
AA866031.1 | cyclophilin | Entamoeba histolytica | details
AAC24632.1 | cyclophilin type peptidyl-prolyl cis-trans isomerase, putative | Leishmania major | details
AAC41300.1 | cyclophilin | Plasmodium falciparum | details
AAC45975.1 | cyclophilin | Plasmodium falciparum | details
AAD55769.1 | cyclophilin | Plasmodium falciparum | details
AAC5985.1 | cyclophilin A | Trypanosoma cruzi | details
AAC59247.1 | cyclophilin | Trypanosoma cruzi | details
AAC5072.1 | cyclophilin | Trypanosoma cruzi | details
AAM21054.1 | cyclophilin | Entamoeba histolytica | details
AAN35754.1 | cyclophilin, putative | Plasmodium falciparum | details
AAN35113.1 | cyclophilin, putative | Plasmodium falciparum | details
AAN35236.1 | cyclophilin | Plasmodium falciparum | details
AAN35836.1 | cyclophilin, putative | Plasmodium falciparum | details
AZ14584.1 | peptidyl-prolyl cis-trans isomerase (cyclophilin), putative | Leishmania major | details
AZ14699.1 | peptidyl-prolyl cis-trans isomerase (cyclophilin-40), putative | Leishmania major | details

### Orthologs by OrthoMCL [go to paper]

**Group**: cyclophilin

### Protozoa EST Similarity (Top 10 Blastx)

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Score</th>
<th>eValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>CX086982</td>
<td>EHA351TR E. histolytica Normalized cDNA library Entamoeba histolytica cDNA, mRNA sequence.</td>
<td>350.0</td>
<td>6e-97</td>
</tr>
<tr>
<td>CX097413</td>
<td>EHAH645TR E. histolytica Normalized cDNA library Entamoeba histolytica cDNA, mRNA sequence.</td>
<td>350.0</td>
<td>7e-97</td>
</tr>
<tr>
<td>CX089961</td>
<td>EHAE652TR E. histolytica Normalized cDNA library Entamoeba histolytica cDNA, mRNA sequence.</td>
<td>347.0</td>
<td>5e-96</td>
</tr>
<tr>
<td>CX096169</td>
<td>EHA017TR E. histolytica Normalized cDNA library Entamoeba histolytica cDNA, mRNA sequence.</td>
<td>338.0</td>
<td>2e-93</td>
</tr>
<tr>
<td>CX082315</td>
<td>EHA072TR E. histolytica Normalized cDNA library Entamoeba histolytica cDNA, mRNA sequence.</td>
<td>336.0</td>
<td>8e-93</td>
</tr>
</tbody>
</table>

---

Integrated System for genotyping Parasitic Protozoa based on Universal orthologous genes

Desing primer for genotyping Parasitic Protozoa Based on Universal orthologous genes

Base Calling chromato files and characterize sequences using Blast

Do Phylogenetic Analyses using: Arpa
or
Evolutionary Analyses using: Pedro

OR

PEDRO
Methodology

DNA extraction

PCR using 33 pair of designed primers

Agarose gel visualization

PCR product purification

Sequencing

Annotation

Phylogenetic analyses
Forward primer (COG0092F) in the multiple alignment

Reverse primer (COG0092F) in the multiple alignment
Genotyping system visualization (multiple alignment) using Jalview
Phylogenetic tree inferred with Maximum Likelihood for KOG0400 (Ribosomal protein S15P/S13E). 1000 replications (bootstrap). Sequences from genbank and generated with our primers are shown.
Methodology: short version

380 sequences from 36 UO → 220 sequences validated, corresponding to 20 UO

33 pair of primers designed

8 pair of primers with good potential for multiplex PCR

Sequencing of 4 genes

Genotyping:
- 2 genes showed good potential/variability
- 2 genes showed low variability
OrthoSearch Workflow (Orthologs inference)

COG/KOG DB → MAFFT → hmmbuild

HMMER

hmmcalibrate → hmmsearch

Ptn DB → hmpfam

Reciprocals Best Hits Finder

formatdb → fastacmd

BLAST

Reannotated sequences → InterPRO
### OrthoSearch results

<table>
<thead>
<tr>
<th>Parasite</th>
<th>Total Hits</th>
<th>Confirmed annotation (Interpro)</th>
<th>Partial annotations (re)annotated</th>
<th>Hypothetical protein (re)annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td>T. cruzi</td>
<td>147</td>
<td>141</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>T. brucei</td>
<td>135</td>
<td>129</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>L. major</td>
<td>165</td>
<td>157</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>P. falciparum</td>
<td>111</td>
<td>107</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>E. histolytica</td>
<td>98</td>
<td>90</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
Protozoan Orthologs
(OrthoMCL-based)

* Some can be protozoa-specific

BioEDDB Consortium
<table>
<thead>
<tr>
<th></th>
<th>PRIMER 19</th>
<th>PRIMER 20</th>
<th>PRIMER 23</th>
<th>PRIMER 143</th>
<th>PRIMER 172</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. mexicana</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>L. braziliensis</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>L. major</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>L. chagasi</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L. amazonensis</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>L. guayanensis</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. cruzi</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. brucei</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. evansi</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. vivax</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. congolense</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Home Recent Results

Choose program to use and database to search:

Program: blastn  Database: refseq_genomic

Enter sequence below in FASTA format

Optional parameters

Threshold  Expect  Matrix  Filter  Descriptions  Alignments  Hits  Gapped
0  10  BLOSUM-62  default  500  250  0  yes

Alignment View Option

pairwise

21883
Search type: DNA sequences

Enter with the sequence in the format FASTA:

Or select multifasta file: Choose File - no file selected

nucleotide

PHYML: Yes

Evolutionary model: HKY

Bootstrap [10 - 1000]: 100

Substitution Rate Categories: 4
The **System for Integrating Genomic Resources and Analyses (STINGRAY)** is a user-friendly web-based system designed to analyse genomic data in the context of a pipeline. It was developed for Linux systems using Perl, Bioperl, CGI, Apache and MySQL. EST and GSS data can especially benefit using the system since it can accept:

a) chromatograms,
b) download of sequences from GenBank,
c) FASTA files stored locally or
d) a combination of all 3

STINGRAY uses the phred/phrap package to process chromatograms, evaluate quality of traces and remove vector contamination. Clusterisation is done using the CAP3 program. Sequences are submitted to Blast analyses with database chosen by the user.

Conserved domain searches are also performed using the CDD tool of NCBI. Interpro have been implemented in the pipeline as well, and tRNA-Scan, Fastt, SignalP and OrthoMCL are available too. Beside that, it is possible to find genes or ORFs using GLIMMER, GlimmerHMM and by Orthologous Groups approach (ORF by COGes), and then search the analysis mentioned above. Once all the above similarities searches have been performed, the user can select some contigs and results for multiple alignment using ClustalW and phylogeny analyses using Phylip. STINGRAY offers flexibility to the users by allowing them to configure the parameters of the used programs. All the results can be filtered according to algorithms and database used, facilitating user visualization. Furthermore, statistics and graphical results of all the analysis and processed chromatograms are presented.
Collaborators

Marta Mattoso
M. Luiza Machado Campos
Edmundo C. Grisard
“Yoko” Cavalcanti

Students

Sérgio Serra
Kary Soriano Ocaña
Fabio Coutinho
Edgar Sarmiento
Diogo Tschoeke
Rafael Cuadrat
Daniel Loureiro
Rodrigo Jardim
Luisa Pitaluga

Recent collaborators:

Elisa Cupolillo
Claudia D. Levy

Funding

CNPq
FINEP
CAPES
FAPERJ
European Union (EELA-2)