D326–D328 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh112

ApiEST-DB: analyzing clustered EST data of the apicomplexan parasites

Li Li, Jonathan Crabtree[1], Steve Fischer[1], Deborah Pinney[1], Christian J. Stoeckert Jr[1], L. David Sibley[2] and David S. Roos*

Department of Biology and [1]Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, USA and [2]Department of Molecular Microbiology, Washington University School of Medicine, St Louis, MO 63108, USA

Received August 15, 2003; Revised and Accepted October 14, 2003

ABSTRACT

ApiEST-DB (http://www.cbil.upenn.edu/paradbs-servlet/) provides integrated access to publicly available EST data from protozoan parasites in the phylum Apicomplexa. The database currently incorporates a total of nearly 100 000 ESTs from several parasite species of clinical and/or veterinary interest, including *Eimeria tenella*, *Neospora caninum*, *Plasmodium falciparum*, *Sarcocystis neurona* and *Toxoplasma gondii*. To facilitate analysis of these data, EST sequences were clustered and assembled to form consensus sequences for each organism, and these assemblies were then subjected to automated annotation via similarity searches against protein and domain databases. The underlying relational database infrastructure, Genomics Unified Schema (GUS), enables complex biologically based queries, facilitating validation of gene models, identification of alternative splicing, detection of single nucleotide polymorphisms, identification of stage-specific genes and recognition of phylogenetically conserved and phylogenetically restricted sequences.

INTRODUCTION

The protozoan phylum Apicomplexa separated from other eukaryotic lineages prior to the divergence of animals and fungi. Apicomplexa are believed to be most closely related to the dinoflagellates and ciliates, which together form the alveolates. The phylum Apicomplexa includes >5000 named species, all of which are parasites (1–3). Apicomplexan parasites infect a wide range of vertebrate hosts and cause many important diseases in humans and domestic animals. Parasites in the genus *Plasmodium* are responsible for human malaria; *Toxoplasma gondii* and *Cryptosporidium parvum* are widespread opportunistic pathogens that cause disease in immunocompromised patients; *T.gondii* is a prominent source of congenital neurological infection in humans and sheep, while *Neospora caninum* causes fetal abortions in cattle; *Eimeria* species cause coccidiosis in poultry; *Theileria parva* and *Babesia bovis* are important cattle pathogens.

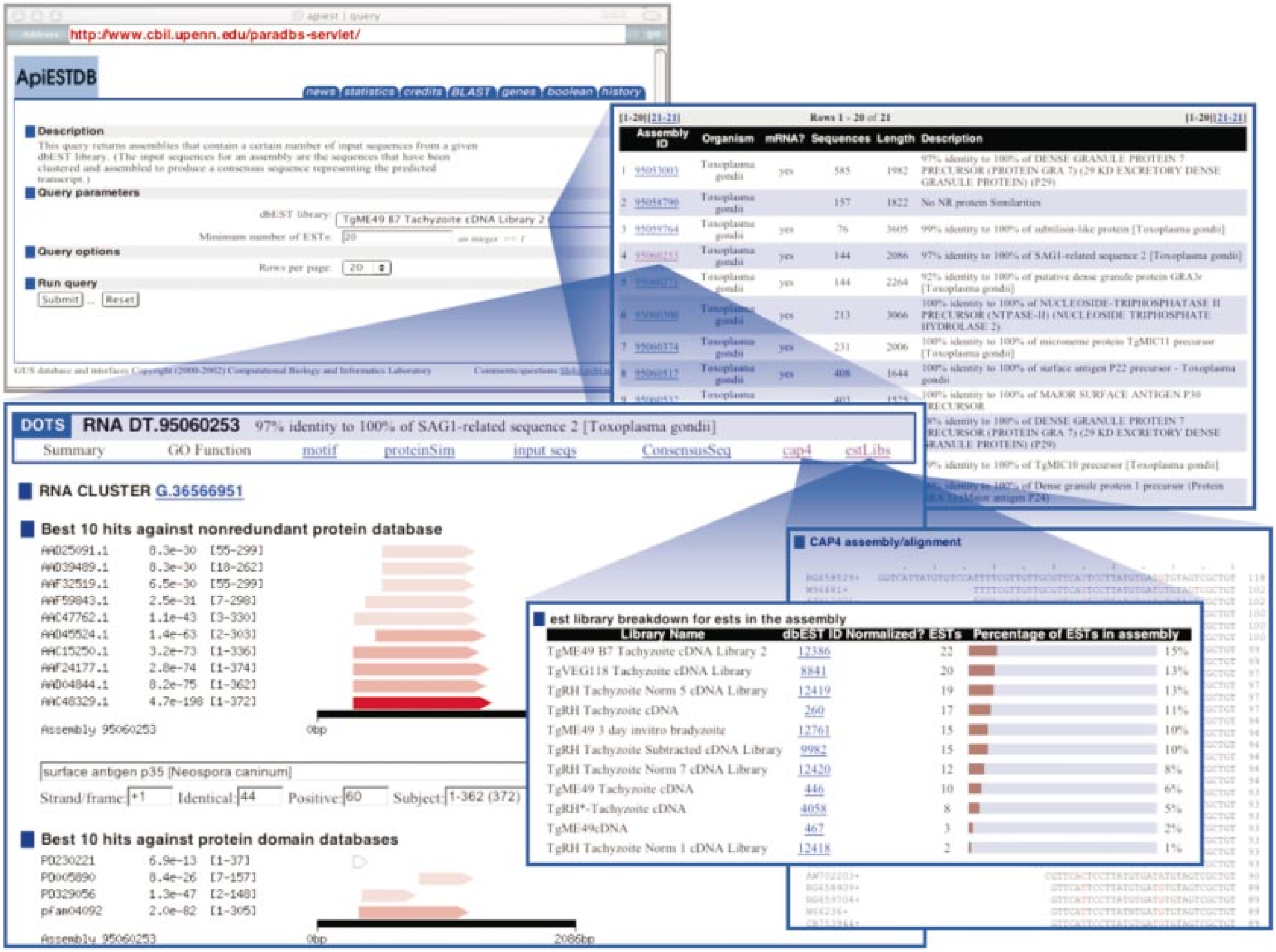
Relatively large EST data sets have been generated for *T.gondii*, *Eimeria tenella* and *Plasmodium falciparum*, and substantial numbers of ESTs are also available for *N.caninum*, *Plasmodium berghei*, *Plasmodium yoelii*, *Sarcocystis neurona* and *T.parva*—with additional sequences for these and other apicomplexan parasites expected in the near future. EST sequencing projects have permitted cost-effective data mining of expressed genes in these parasites prior to sequencing of the full genome. This is particularly true for pathogenic organisms, where abundant transcripts are of interest as likely targets for drug, vaccine and diagnostic development. Even when a complete genome sequence (as for *P.falciparum*) or draft sequence (as for *T.gondii* and *P.yoelii*) are available, EST data provides valuable complementary information, confirming expression and defining intron structure. However, the fragmented and error-prone nature of these sequences, together with their high redundancy and lack of extensive annotation, poses major challenges for utilizing EST data. The goal of ApiEST-DB is to integrate data emerging from various EST sequencing projects, and provide automated annotation and data-mining tools enabling biologically interesting queries from the Apicomplexa research community and beyond.

DATA AND QUERY TOOLS

The current release of ApiEST-DB incorporates 61 116 ESTs from *T.gondii*, 10 023 from *P.falciparum*, 28 550 from *E.tenella*, 4949 from *S.neurona* and 3121 from *N.caninum.* All EST data were downloaded from the NCBI dbEST database (4) (http://www.ncbi.nlm.nih.gov/dbEST/) and loaded into the Genomics Unified Schema (GUS; http://www.gusdb.org) relational database for further analysis, using database and interface tools developed for the Allgenes database (http://www.allgenes.org). To handle redundancy and possible sequencing errors, apicomplexan EST sequences and additional mRNA/cDNA sequences from GenBank (5) were first 'cleaned' by detecting and removing vector sequences, poly(A/T)s and poor-quality ends. Although the obligate intracellular lifestyle of these parasites raises the possibility of contaminating host cell mRNA as a possible concern, we have previously demonstrated that it is generally rare (6).

For each individual species, sequences longer than 50 nt were then clustered by running an 'all-against-all' BLASTN comparison (7) (http://blast.wustl.edu). Species-specific

*To whom correspondence should be addressed. Tel: +1 215 898 2118; Fax: +1 215 898 8780; Email: droos@sas.upenn.edu



apiest | query

http://www.cbil.upenn.edu/paradbs-servlet/

ApiESTDB

Description

This query returns assemblies that contain a certain number of input sequences from a given dbEST library. (The input sequences for an assembly are the sequences that have been clustered and assembled to produce a consensus sequence representing the predicted transcript.)

Query parameters

dbEST library: TgME49 B7 Tachyzoite cDNA Library 2

Minimum number of ESTs: 20 an integer >= 1

Query options

Rows per page: 20

Run query

GUS database and interfaces Copyright (2000-2002) Computational Biology and Informatics Laboratory Comments/questions

[1-20][21-21] Rows 1 - 20 of 21 [1-20][21-21]

	Assembly ID	Organism	mRNA?	Sequences	Length	Description
1	95053003	Toxoplasma gondii	yes	585	1982	97% identity to 100% of DENSE GRANULE PROTEIN 7 PRECURSOR (PROTEIN GRA 7) (29 KD EXCRETORY DENSE GRANULE PROTEIN) (P29)
2	95058790	Toxoplasma gondii		157	1822	No NR protein Similarities
3	95059764	Toxoplasma gondii	yes	76	3605	99% identity to 100% of subtilisin-like protein [Toxoplasma gondii]
4	95060253	Toxoplasma gondii	yes	144	2086	97% identity to 100% of SAG1-related sequence 2 [Toxoplasma gondii]
5	95060271	Toxoplasma gondii	yes	144	2264	92% identity to 100% of putative dense granule protein GRA3r [Toxoplasma gondii]
6	95060300	Toxoplasma gondii	yes	213	3066	100% identity to 100% of NUCLEOSIDE-TRIPHOSPHATASE II PRECURSOR (NTPASE-II) (NUCLEOSIDE TRIPHOSPHATE HYDROLASE 2)
7	95060374	Toxoplasma gondii	yes	231	2006	100% identity to 100% of microneme protein TgMIC11 precursor [Toxoplasma gondii]
8	95060517	Toxoplasma gondii	yes	408	1644	100% identity to 100% of surface antigen P22 precursor - Toxoplasma gondii
9	95060532	Toxoplasma		401	1325	100% identity to 100% of MAJOR SURFACE ANTIGEN P30 RECURSOR
						% identity to 100% of DENSE GRANULE PROTEIN 7 RECURSOR (PROTEIN GRA 7) (29 KD EXCRETORY DENSE RANULE PROTEIN) (P29)
						% identity to 100% of TgMIC10 precursor [Toxoplasma gondii]
						% identity to 100% of Dense granule protein 1 precursor (Protein GRA 1) (Major antigen P24)

DOTS **RNA DT.95060253** 97% identity to 100% of SAG1-related sequence 2 [Toxoplasma gondii]

RNA CLUSTER G.36566951

Best 10 hits against nonredundant protein database

```
AAD25091.1     8.3e-30   [55-299]
AAD39489.1     8.3e-30   [18-262]
AAF32519.1     6.5e-30   [55-299]
AAF59843.1     2.5e-31   [7-298]
AAC47762.1     1.1e-43   [3-330]
AAD45524.1     1.4e-63   [2-303]
AAC15250.1     3.2e-73   [1-336]
AAF24177.1     2.8e-74   [1-374]
AAD04844.1     8.2e-75   [1-362]
AAC48329.1     4.7e-198  [1-372]

Assembly 95060253
```

surface antigen p35 [Neospora caninum]

Strand/frame: +1 Identical: 44 Positive: 60 Subject: 1-362 (372)

Best 10 hits against protein domain databases

```
PD230221      6.9e-13   [1-37]
PD005890      8.4e-26   [7-157]
PD329056      1.3e-47   [2-148]
pfam04092     2.0e-82   [1-305]

Assembly 95060253
```

CAP4 assembly/alignment

est library breakdown for ests in the assembly

Library Name	dbEST ID	Normalized?	ESTs	Percentage of ESTs in assembly
TgME49 B7 Tachyzoite cDNA Library 2	12386		22	15%
TgVEG118 Tachyzoite cDNA Library	8841		20	13%
TgRH Tachyzoite Norm 5 cDNA Library	12419		19	13%
TgRH Tachyzoite cDNA	260		17	11%
TgME49 3 day invitro bradyzoite	12761		15	10%
TgRH Tachyzoite Subtracted cDNA Library	9982		15	10%
TgRH Tachyzoite Norm 7 cDNA Library	12420		12	8%
TgME49 Tachyzoite cDNA	446		10	6%
TgRH*-Tachyzoite cDNA	4058		8	5%
TgME49cDNA	467		3	2%
TgRH Tachyzoite Norm 1 cDNA Library	12418		2	1%

Figure 1. Graphical illustration of a sample query session. A query for *Toxoplasma* assemblies that contain at least 20 ESTs from the library 'TgME49 B7 Tachyzoite cDNA Library 2' (top left) identifies 21 assemblies (top right). The summary page for assembly 95060253 (bottom left) provides a variety of information, including a graphical representation of the top 10 BLAST similarities to protein and domain databases, and links to these results in tabular format, the constituent ESTs contained in the assembly ('input seqs', hot-linked to dbEST), the EST sequence alignment ('cap4'; bottom right, with potential SNPs highlighted in red), and a graphical view of information on library distribution ('estLibs'; bottom center).

sequence clusters were formed by a connected component analysis of all the BLASTN matches with minimum cut-off values of 92% identity over 40 nt length, and without overhanging ends. Clusters were then assembled to form consensus sequences using the CAP4 algorithm (Paracel Inc.; http://www.paracel.com), yielding 18 805 assemblies for *T.gondii*, 5800 for *P.falciparum*, 7089 for *E.tenella*, 1445 for *S.neurona* and 1388 for *N.caninum*. Higher-level relationships were identified based on BLAST similarities among assemblies from a given species, generating 'RNA clusters' that highlight potential gene families, alternative transcripts and differential splicing. Highly represented assemblies for each organism can be searched via a query specifying a minimum number of constituent sequences. Databases of the consensus sequences for each organism are available for BLAST searches with user-defined protein or DNA sequences.

To facilitate gene discovery, ApiEST-DB provides automated annotations of the mRNA assemblies using similarity searches. The consensus sequence for each assembly was searched against the NCBI non-redundant protein database using BLASTX. The best match with P value <1e–10 was used to assign a putative description for each assembly, and a word index was created for these annotations. The results of all BLAST searches with a cut-off value of P <1e–5 were also stored, and an index was created for words appearing in the description of these hits. Finally, we conducted searches against protein domain databases, including ProDom (8) (using BLASTX) and CDD (9) (using RPS-BLAST). Users can conduct keyword searches for assemblies whose putative description or BLAST hits (similar proteins or functional domains) contain any keyword of interest, and can then view a summary of all BLAST hits that were generated by the assembly, with direct links to the external protein or domain databases provided for each subject discovered by the search.

These tools are comparable to several previously reported software pipelines designed for annotating EST sequences (10–12), although ApiEST-DB does not provide a mechanism for clustering of EST data sets submitted by the user. The strength of ApiEST-DB derives from its sophisticated relational infrastructure which provides a platform for extensive data integration and cross-species comparison. The web interface also facilitates functional analyses such as stage-specific expression and SNP identification.

In addition to assisting in gene identification and annotation, one interesting class of questions that can be examined using EST data relates to strain or developmental stage specificity. To enable such analysis, information on each library from which ESTs have been generated is stored, allowing library distribution to be displayed for each mRNA assembly, and enabling queries based on the number of ESTs from a given library. The database also supports Boolean combinations of the various queries available, and retains a history of searches performed in each session, enabling results to be selectively combined by users.

For each query result, users are directed to a summary page providing links to detailed information about the assembly, including the best 10 BLAST hits to protein sequences and protein domain databases (in both tabular and graphical form), the CAP4 alignment of constituent sequences (with potential SNPs highlighted in red) and the breakdown of ESTs in each assembly according to their library source (facilitating identification of strain or stage specificity). Summary tables are also provided for cross-comparisons between apicomplexan parasites, highlighting gene families that are specific to the Apicomplexa. Such gene families are likely to be useful targets for the development of broad-spectrum vaccines, drugs and/or diagnostic reagents.

FUTURE DIRECTIONS

With the further expansion of EST and genome projects associated with the phylum Apicomplexa and other related taxa (13), ApiEST-DB will incorporate this new information, including sequence data from additional species. New builds of ApiEST-DB will be coordinated with the release of new sequences generated from apicomplexan EST projects. ApiEST-DB uses the same underlying database infrastructure as PlasmoDB (14) (http://PlasmoDB.org), ToxoDB (15) and CryptoDB (http://CryptoDB.org) (16), allowing data integration with genomic data and other data types related to gene expression, including microarray, SAGE tag and proteomics data from *Plasmodium* spp. and *T.gondii.* As more species are included in ApiEST-DB, more sophisticated queries and tools for cross-species comparison and phylogenetic analysis will also be provided.

ACKNOWLEDGEMENTS

We thank members of the Roos and Stoeckert groups who provided useful advice on the development and utility of the ApiEST-DB resource, including Brian Brunk, Bindu Gajria and Jessica Kissinger. We are grateful to the sequencing centers at Washington University and elsewhere that have produced apicomplexan EST sequences, and the many individual investigators who have made available the reagents and information on which this database depends. This project was supported by grants from the NIH, Merck Research Laboratories, USDA and the Burroughs Wellcome Fund. D.S.R. is a Senior Scholar in Global Infectious Diseases of the Ellison Medical Foundation.

REFERENCES

1. Cavalier-Smith,T. (1993) Kingdom protozoa and its 18 phyla. *Microbiol. Rev.*, **57**, 953–994.
2. Levine,N.D. (1988) Progress in taxonomy of the Apicomplexan protozoa. *J. Protozool.*, **35**, 518–520.
3. Vivier,E. and Desportes,I. (1989) *Apicomplexa*. In Margulis,L., Corliss,J., Melkonian,M. and Chapman,D. (eds), *Handbook of Protoctista*. Jones and Bartlett, Boston, MA.
4. Rodriguez-Tome,P. (1997) Searching the dbEST database. *Methods Mol. Biol.*, **69**, 269–283.
5. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
6. Li,L., Brunk,B.P., Kissinger,J.C., Pape,D., Tang,K., Cole,R.H., Martin,J., Wylie,T., Dante,M., Fogarty,S.J. *et al.* (2003) Gene discovery in the Apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. *Genome Res.*, **13**, 443–454.
7. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
8. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief. Bioinform.*, **3**, 246–251.
9. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
10. Ayoubi,P., Jin,X., Leite,S., Liu,X., Martajaja,J., Abduraham,A., Wan,Q., Yan,W., Misawa,E. and Prade,R.A. (2002) PipeOnline 2.0: automated EST processing and functional data sorting. *Nucleic Acids Res.*, **30**, 4761–4769.
11. Hotz-Wagenblatt,A., Hankeln,T., Ernst,P., Glatting,K.H., Schmidt,E.R. and Suhai,S. (2003) ESTAnnotator: a tool for high throughput EST annotation. *Nucleic Acids Res.*, **31**, 3716–3719.
12. Mao,C., Cushman,J.C., May,G.D. and Weller,J.W. (2003) ESTAP—an automated system for the analysis of EST data. *Bioinformatics*, **19**, 1720–1722.
13. Tarleton,R.L. and Kissinger,J. (2001) Parasite genomics: current status and future prospects. *Curr. Opin. Immunol.*, **13**, 395–402.
14. Bahl,A., Brunk,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C., Labo,P. *et al.* (2003) PlasmoDB: the *Plasmodium* genome resource: a database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.
15. Kissinger,J.C., Gajria,B., Li,L., Paulsen,I.T. and Roos,D.S. (2003) ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res.*, **31**, 234–236.
16. Puiu,D., Enomoto,S., Buck,G.A., Abrahamsen,M.S. and Kissinger,J.C. (2004) CryptoDB: the *Cryptosporidium* genome resource. *Nucleic Acids Res.*, **32**, D329–D331.