Nucleic Acids Research, 2006, Vol. 34, Database issue ***D419–D422***
doi:10.1093/nar/gkj078

CryptoDB: a *Cryptosporidium* bioinformatics resource update

Mark Heiges, Haiming Wang, Edward Robinson, Cristina Aurrecoechea, Xin Gao[1,2], Nivedita Kaluskar[1], Philippa Rhodes[1], Sammy Wang[1], Cong-Zhou He[1], Yanqi Su[1], John Miller[1], Eileen Kraemer[1] and Jessica C. Kissinger[2,]*

Center for Tropical and Emerging Global Diseases, [1]Department of Computer Science and [2]Department of Genetics, University of Georgia, Athens, GA, USA

Received September 15, 2005; Accepted October 11, 2005

ABSTRACT

The database, CryptoDB (http://CryptoDB.org), is a community bioinformatics resource for the AIDS-related apicomplexan-parasite, *Cryptosporidium*. CryptoDB integrates whole genome sequence and annotation with expressed sequence tag and genome survey sequence data and provides supplemental bioinformatics analyses and data-mining tools. A simple, yet comprehensive web interface is available for mining and visualizing the data. CryptoDB is allied with the databases PlasmoDB and ToxoDB via ApiDB, an NIH/NIAID-fundedBioinformatics Resource Center. Recent updates to CryptoDB include the deposition of annotated genome sequences for *Cryptosporidium parvum* and *Cryptosporidium hominis*, migration to a relational database (GUS), a new query and visualization interface and the introduction of Web services.

INTRODUCTION

The Apicomplexan parasite *Cryptosporidium* is a global causative agent of severe and chronic diarrheal disease in humans and other animals. As no reliable chemo- or immuno-therapy is currently available, infections can be life threatening for people with a compromised immune system, such as AIDS patients. The pathogen is typically spread via contaminated drinking water and is resistant to water chlorination and filtration (1). Because of the water safety threat to public health, *Cryptosporidium* is ranked as a Category B Biodefense Pathogen by the National Institutes of Health. Bioinformatics analysis plays an important role in understanding the biology of and identification of potential drug targets in this medically important parasite. To aid the research community in this line of inquiry, the online database CryptoDB continues to update and expand its role of warehousing and interfacing *Cryptosporidium* genome sequence, annotation, sequence analysis and other *Cryptosporidium*-related information.

UPDATED DATASET

Version 3.0 of CryptoDB was released in April 2005 and contains the published genome sequence and annotation for *Cryptosporidium hominis*, strain TU502 (2) and *Cryptosporidium parvum*, strain IOWA (3). The database houses copies of assembled genome contigs and gene annotations deposited in GenBank (4) by the sequence generators. The *C.parvum* genome sequence is represented by 18 contigs ranging in size from 17 kb to 1.2 Mb in length and annotated with 3885 total genes. The *C.hominis* genome sequence is represented in 1422 contigs ranging in size from a few hundred to 90 thousand base pairs in length and annotated with 3956 total genes. *C.parvum* chromosome 6 has been independently sequenced and annotated (5) and is represented in the database.

In addition to the data provided by genome sequencing efforts, ∼6 Mb of genome survey sequence (GSS) and expressed sequence tag (EST) (6) data are incorporated. The ESTs are clustered into RNA transcripts and aligned to the genome using the methodology applied at ApiEST-DB (7).

Gene annotations provided by the genome sequencing centers are augmented with supplemental analyses. Pre-computed BLASTX analyses of *Cryptosporidium* contigs versus the GenBank non-redundant protein database and EST alignments to contig sequences offer supporting evidence for gene predictions. Potential syntenic relationships of the *C.hominis* and *C.parvum* contig sequences are calculated and graphically displayed. Protein feature predictions of signal peptides and transmembrane domains are provided. Open reading frames >50 and >100 amino acids in length have been calculated for all nucleic acid sequences in all six reading frames.

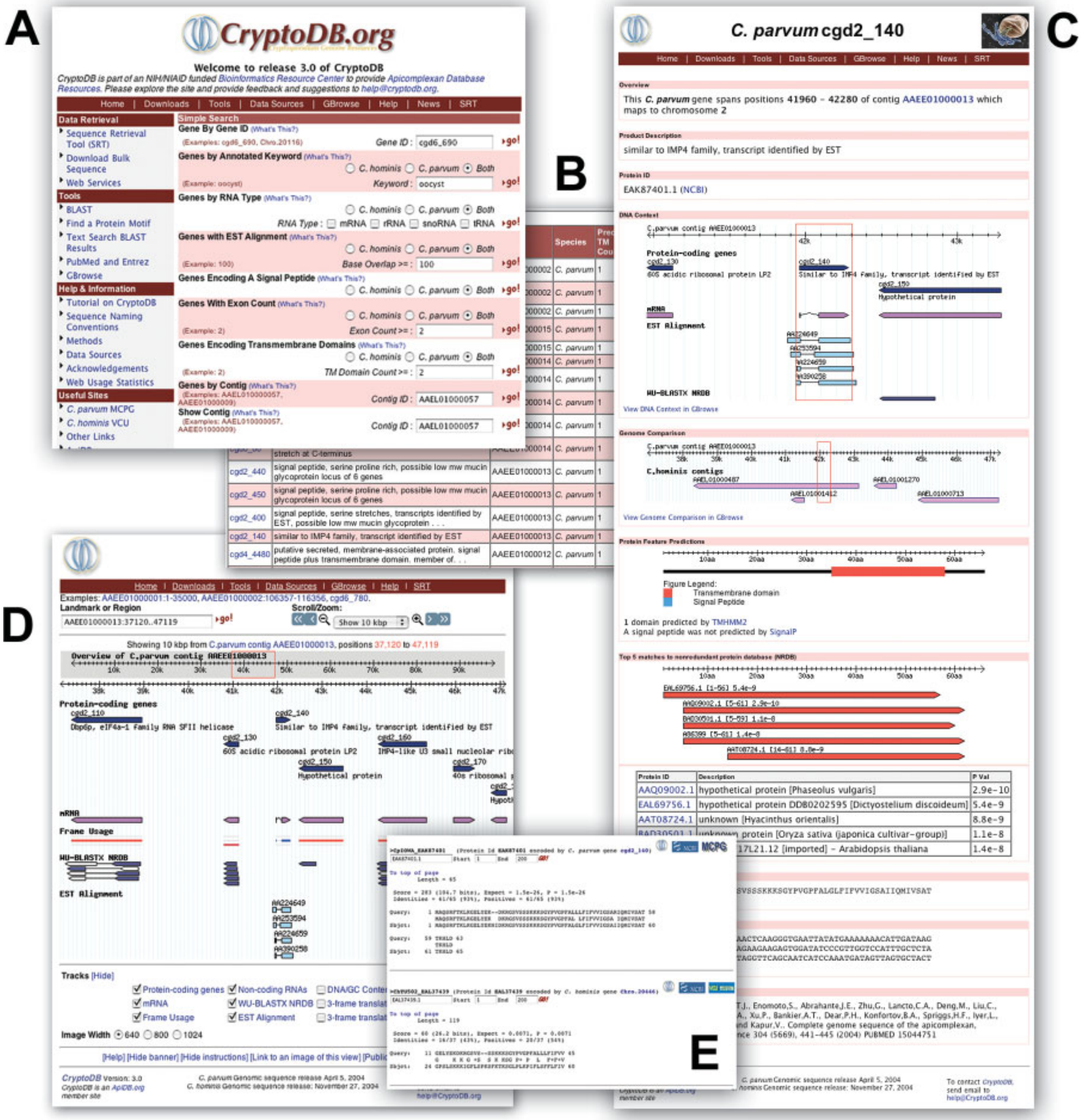
*To whom correspondence should be addressed. Tel: +1 706 542 6562; Fax: +1 706 542 3910; Email: jkissing@uga.edu

All sequence datasets are available for bulk download in FASTA format. Programmatic access to selected resources is provided via Web service interfaces.

IMPROVED DATABASE AND WEB INTERFACE

CryptoDB 3.0 is backed by a relational database utilizing the Genomics Unified Schema (GUS, GUSdb.org) (8) and Oracle 10g. Migration to a relational database architecture marks a major improvement over previous releases because of the new services and resources that can now be offered.

The CryptoDB web interface provides a set of forms through which users can easily query the annotation and pre-computed analysis data (Figure 1A). Queries for contig sequence, gene and protein features are possible and can be restricted to either or both of the hosted species genomes. At the gene level, users can conduct text searches of gene product descriptions, search for genes by RNA type (mRNA, rRNA, snoRNA and tRNA) and find genes having alignments to *C.parvum* ESTs. For protein features, users can select genes predicted by SignalP (9) to encode a signal peptide or predicted by TMHMM (10) to contain transmembrane domains. Users may also retrieve a specific gene by locus tag or a contig



Welcome to release 3.0 of CryptoDB

CryptoDB is part of an NIH/NIAID funded Bioinformatics Resource Center to provide Apicomplexan Database Resources. Please explore the site and provide feedback and suggestions to help@cryptodb.org.

Simple Search

Gene By Gene ID (What's This?) (Examples: cgd6_690, Chro.20116) *Gene ID*: cgd6_690

Genes by Annotated Keyword (What's This?) ○ *C. hominis* ○ *C. parvum* ● *Both* (Example: oocyst) *Keyword*: oocyst

Genes by RNA Type (What's This?) ○ *C. hominis* ○ *C. parvum* ● *Both* *RNA Type*: □ mRNA □ rRNA □ snoRNA □ tRNA

Genes with EST Alignment (What's This?) ○ *C. hominis* ○ *C. parvum* ● *Both* (Example: 100) *Base Overlap >=*: 100

Genes Encoding A Signal Peptide (What's This?) ○ *C. hominis* ○ *C. parvum* ● *Both*

Genes With Exon Count (What's This?) ○ *C. hominis* ○ *C. parvum* ● *Both* (Example: 2) *Exon Count >=*: 2

Genes Encoding Transmembrane Domains (What's This?) ○ *C. hominis* ○ *C. parvum* ● *Both* (Example: 2) *TM Domain Count >=*: 2

Genes by Contig (What's This?) (Examples: AAEL01000057, AAEE01000009) *Contig ID*: AAEL01000057

Show Contig (What's This?) (Examples: AAEL01000057, AAEE01000009) *Contig ID*: AAEL01000057

			Species	Pred TM Cou
		000002	C. parvum	1
		000002	C. parvum	1
		000002	C. parvum	1
		000015	C. parvum	1
		000015	C. parvum	1
		000014	C. parvum	1
		000014	C. parvum	1
		000014	C. parvum	1
		000014	C. parvum	1
	stretch at C-terminus	000014	C. parvum	1
cgd2_440	signal peptide, serine proline rich, possible low mw mucin glycoprotein locus of 6 genes	AAEE01000013	C. parvum	1
cgd2_450	signal peptide, serine proline rich, possible low mw mucin glycoprotein locus of 6 genes	AAEE01000013	C. parvum	1
cgd2_400	signal peptide, serine stretches, transcripts identified by EST, possible low mw mucin glycoprotein . . .	AAEE01000013	C. parvum	1
cgd2_140	similar to IMP4 family, transcript identified by EST	AAEE01000013	C. parvum	1
cgd4_4480	putative secreted, membrane-associated protein, signal peptide plus transmembrane domain, member of. . .	AAEE01000012	C. parvum	1

C. parvum cgd2_140

Overview: This *C. parvum* gene spans positions 41960 - 42280 of contig AAEE01000013 which maps to chromosome 2

Product Description: similar to IMP4 family, transcript identified by EST

Protein ID: EAK87401.1 (NCBI)

DNA Context: C.parvum contig AAEE01000013; cgd2_130 60S acidic ribosomal protein LP2; cgd2_140 Similar to IMP4 family, transcript identified by EST; cgd2_150 Hypothetical protein; AA224649, AA253594, AA224659, AA390258. View DNA Context in GBrowse

Genome Comparison: C.parvum contig AAEE01000013; C.hominis contigs AAEL01000487, AAEL01001270, AAEL01001412, AAEL01000713. View Genome Comparison in GBrowse

Protein Feature Predictions

Figure Legend:
🟥 Transmembrane domain
🟦 Signal Peptide

1 domain predicted by TMHMM2
A signal peptide was not predicted by SignalP

Top 5 matches to nonredundant protein database (NRDB): EAL69756.1 [1-56] 5.4e-9; AAQ09002.1 [5-61] 2.9e-10; BAD30501.1 [5-59] 1.1e-8; A86399 [5-61] 1.4e-8; AAT08724.1 [16-61] 8.8e-9

Protein ID	Description	P Val
AAQ09002.1	hypothetical protein [Phaseolus vulgaris]	2.9e-10
EAL69756.1	hypothetical protein DDB0202595 [Dictyostelium discoideum]	5.4e-9
AAT08724.1	unknown [Hyacinthus orientalis]	8.8e-9
BAD30501.1	unknown protein [Oryza sativa (japonica cultivar-group)]	1.1e-8
	17L21.12 [imported] - Arabidopsis thaliana	1.4e-8

SVSSSKKKSGYPVGPFALGLFIFVVIGSAIIQMIVSAT

AACTCAAGGGTGAATTATATGAAAAAAACATTGATAAG
AGAAGAAGAGTGGATATCCCGTTGGTCCATTTGCTCTA
TAGGTTCAGCAATCATCCAAATGATAGTTAGTGCTACT

T.J., Enomoto,S., Abrahante,J.E., Zhu,G., Lancto,C.A., Deng,M., Liu,C., A., Xu,P., Bankier,A.T., Dear,P.H., Konfortov,B.A., Spriggs,H.F., Iyer,L., and Kapur,V.. Complete genome sequence of the apicomplexan, nce 304 (5669), 441-445 (2004) PUBMED 15044751

C. parvum Genomic sequence release April 5, 2004
C. hominis Genomic sequence release: November 27, 2004
To contact CryptoDB, send email to help@CryptoDB.org

Home | Downloads | Tools | Data Sources | GBrowse | Help | SRT

Examples: AAEE01000001:1-35000, AAEE01000002:106357-116356, cgd6_780.

Landmark or Region: AAEE01000013:37120..47119 Scroll/Zoom: Show 10 kbp

Showing 10 kbp from C.parvum contig AAEE01000013, positions 37,120 to 47,119

Overview of C.parvum contig AAEE01000013

cgd2_110 Dbp6p, eIF4a-1 family RNA SFII helicase; cgd2_130 60S acidic ribosomal protein LP2; cgd2_140 Similar to IMP4 family, transcript identified by EST; cgd2_150 Hypothetical protein; cgd2_160 IMP4-like U3 small nucleolar ribo; cgd2_170 40s ribosomal; AA224649, AA253594, AA224659, AA390258

Tracks [Hide]
☑ Protein-coding genes ☑ Non-coding RNAs □ DNA/GC Conten
☑ mRNA ☑ WU-BLASTX NRDB □ 3-frame translat
☑ Frame Usage ☑ EST Alignment □ 3-frame translat

Image Width ● 640 ○ 800 ○ 1024

[Help] [Hide banner] [Hide instructions] [Link to an image of this view] [Public

CryptoDB Version: 3.0
CryptoDB is an ApiDB.org member site
C. parvum Genomic sequence release April 5, 2004
C. hominis Genomic sequence release: November 27, 2004
help@CryptoDB.org

```
>CpIOWA_EAK87401   (Protein Id EAK87401 encoded by C. parvum gene cgd2_140)
EAK87401.1    Start 1    End 200
To top of page
        Length = 65

 Score = 283 (104.7 bits), Expect = 1.5e-26, P = 1.5e-26
 Identities = 61/65 (93%), Positives = 61/65 (93%)

>ChUGA02_EAL37439 (Protein Id EAL37439 encoded by C. hominis gene Chro.20446)
EAL37439.1    Start 1    End 200
To top of page
        Length = 119

 Score = 60 (26.2 bits), Expect = 0.0071, P = 0.0071
 Identities = 16/37 (43%), Positives = 20/37 (54%)
```

Figure 1. Database functionality. (**A**) Searches are initiated via queries provided on the web site's front page. (**B**) The results are returned as a summary table with links to detailed record pages. (**C**) Detailed record page with summary of all available data/information for this gene or contig sequence. (**D**) GBrowse of genomic region of interest provides a graphical view of annotations and similarity analyses. (**E**) Search results, such as BLAST reports, are linked to detailed records and to a sequence retrieval utility.

sequence by accession number. *Ad hoc* data selections not obtainable via the provided queries may be requested by email to help@cryptodb.org.

Gene pages and contig sequence pages provide a detailed view of annotation and analysis for a given record in the database. Gene pages contain a text overview of the gene, including the coordinate position on its contig and product description when available (Figure 1C). GBrowse (11) has been utilized to provide a graphic display of annotated gene features and the data mapped to the genome, BLAST hits, ESTs, etc. (Figure 1D).

The web interface includes a mechanism to allow users to readily download the sequences and other attributes associated with their query result set. A query history permits users to track their searches and combine them into more complex queries across data types (e.g. 'list all genes on chromosome 3 that contain transmembrane domains').

ANALYSIS AND RESEARCH TOOLS

Several tools for data mining augment the published annotations and pre-computed analyses. Users may BLAST their own sequences against the genomic contig, annotated protein, GSS and EST sequence databases (Figure 1E). A motif search tool finds protein sequences with PROSITE (12) or user-defined amino acid patterns. Keywords from *Cryptosporidium* genomic sequences versus GenBank NRDB BLASTX results are indexed and searchable. In each case, the results contain links back to detailed gene, protein or contig record pages or to external databases (e.g. GenBank) as appropriate (Figure 1B).

To facilitate tracking of the latest literature, PubCrawler (13) is used to poll NCBI's PubMed and GenBank each week day for new *Cryptosporidium*-related updates.

AFFILIATIONS

CryptoDB is a member of ApiDB.org, an NIH/NIAID funded Bioinformatics Resource Center (BRC) for Biodefense and Emerging or Re-Emerging Infectious Diseases (www.niaid.nih.gov/dmid/genomes/brc/default.htm). Other ApiDB members include the genome databases for *Plasmodium* (PlasmoDB) (14,15) and *Toxoplasma* (ToxoDB) (16) and the Apicomplexan EST database, ApiEST-DB (7). CryptoDB and other member databases are linking to ApiDB in a coordinated effort to promote comparative studies and ease of access across these apicomplexan genomes.

WEB SERVICES AND DATABASE NEWS

To facilitate database integration with ApiDB, other NIAID BRC's and programmatic access of CryptoDB by others, web services for CryptoDB have been implemented. Web services are pieces of software that can communicate across the Internet to build distributed applications. They can do this regardless of the software used for their implementation as long as they use a common protocol, SOAP (17). CryptoDB uses SOAP and provides published WSDL files and sample client software in Java (using Axis) and PERL (using SOAP::Lite). Currently, one service that retrieves FASTA sequence files is active. Additional services and infrastructure (18) are planned.

To facilitate the dissemination of news and updates related to CryptoDB, we have established a Really Simple Syndication news feed (RSS) that is displayed on the home page of CryptoDB and ApiDB and can be read by any RSS news aggregator.

FUTURE PLANS

CryptoDB is fully funded and staffed with biologists and software developers with close ties to software developers for GUS, ToxoDB and PlasmoDB. This fertile ground will support many opportunities for frequent database updates and expansions with new data types, analyses, data-mining tools and visual displays. Gene ontology terms and protein feature signatures from InterProScan (19) analyses will be included with gene records. Improvements to visualization of genome-wide synteny are planned. SRI International's Pathway Tools software (20) is being added to facilitate analyses of metabolic pathways in both annotated *Cryptosporidium* genome sequences. Future releases of CryptoDB will publish this information, as 'CryptoCyc' for querying and visualization in a graphical display.

To facilitate data sharing, CryptoDB has the capacity to activate a Distributed Annotation Server (DAS) (21) via a DAS-GUS adapter if needed.

CryptoDB will continue to work with the ApiDB consortium to further integrate its resources with other apicomplexan genome sites. Database federating technologies, web services and portal designs are currently being implemented toward this end. Data exchange and interoperability with other NIAID Bioinformatics Resource Centers will be a continued effort.

ACKNOWLEDGEMENTS

CryptoDB is one component of the ApiDB project co-administered by David Roos, Christian Stoeckert and Jessica Kissinger. The architecture and infrastructure development applied to CryptoDB has greatly benefited from collaborative ApiDB efforts directed toward database integration, GUS schema development and the new web development kit. The authors thank our database development collaborators: Steve Fischer, Bindu Gajria, Thomas Gan, John Iodice, Aaron Mackey, Jules Milgram, Debbie Pinney and Michael Saffitz. This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400037C. Funding to pay the Open Access publication charges for this article was provided by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400037C.

Conflict of interest statement. None declared.

REFERENCES

1. Fayer,R. (1997) *Cryptosporidium and Cryptosporidiosis*. CRC Press, Inc., Boca Raton, FL.
2. Xu,P., Widmer,G., Wang,Y., Ozaki,L.S., Alves,J.M., Serrano,M.G., Puiu,D., Manque,P., Akiyoshi,D., Mackey,A.J. *et al*. (2004) The genome of *Cryptosporidium hominis*. *Nature*, **431**, 1107–1112.

3. Abrahamsen,M.S., Templeton,T.J., Enomoto,S., Abrahante,J.E., Zhu,G., Lancto,C.A., Deng,M., Liu,C., Widmer,G., Tzipori,S. *et al.* (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, **304**, 441–445.
4. Xu,P., Widmer,G., Wang,Y., Ozaki,L.S., Alves,J.M., Serrano,M.G., Puiu,D., Manque,P., Akiyoshi,D., Mackey,A.J. *et al.* (2004) Corrigendum: the genome of *Cryptosporidium hominis*. *Nature*, **432**, 415.
5. Bankier,A.T., Spriggs,H.F., Fartmann,B., Konfortov,B.A., Madera,M., Vogel,C., Teichmann,S.A., Ivens,A. and Dear,P.H. (2003) Integrated mapping, chromosomal sequencing and sequence analysis of *Cryptosporidium parvum*. *Genome Res.*, **13**, 1787–1799.
6. Strong,W.B. and Nelson,R.G. (2000) Preliminary profile of the *Cryptosporidium parvum* genome: an expressed sequence tag and genome survey sequence analysis. *Mol. Biochem. Parasitol.*, **107**, 1–32.
7. Li,L., Crabtree,J., Fischer,S., Pinney,D., Stoeckert,C.J.Jr, Sibley,L.D. and Roos,D.S. (2004) ApiEST-DB: analyzing clustered EST data of the apicomplexan parasites. *Nucleic Acids Res.*, **32**, D326–D328.
8. Davidson,S.B., Crabtree,J., Brunk,B., Schug,J., Tannen,V., Overton,G.C. and Stoeckert,J.C.J. (2001) K2/Kleisli and GUS: experiments in integrated access to genomic data sources. *IBM systems Journal*, **40**, 512–531.
9. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
10. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
11. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
12. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.*, **3**, 265–274.
13. Hokamp,K. and Wolfe,K.H. (2004) PubCrawler: keeping up comfortably with PubMed and GenBank. *Nucleic Acids Res.*, **32**, W16–W19.
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., Brunk,B.P., Crabtree,J., Fraunholz,M.J., Gajria,B., Milgram,A.J., Pearson,D.S., Schug,J., Bahl,A., Diskin,S.J. *et al.* (2002) The *Plasmodium* genome database. *Nature*, **419**, 490–492.
16. 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.
17. Stein,L. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119–120.
18. Sivashanmugam,K., Miller,J.A., Sheth,A.P. and Verma,K. (2004) Framework for semantic web process composition. *International Journal of Electronic Commerce*, **9**, 71–106.
19. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
20. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18** (Suppl. 1), S225–S232.
21. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.