Comparative Analysis of Apicomplexa and Genomic Diversity in Eukaryotes

Thomas J. Templeton,[1,7,8] Lakshminarayan M. Iyer,[2,7] Vivek Anantharaman,[2] Shinichiro Enomoto,[3] Juan E. Abrahante,[3] G.M. Subramanian,[5] Stephen L. Hoffman,[6] Mitchell S. Abrahamsen,[3,4] and L. Aravind[2,8]

[1]Department of Microbiology and Immunology, Weill Medical College and the Program in Immunology and Microbial Pathogenesis, Weill Graduate School of Medical Sciences of Cornell University, New York, New York 10021, USA; [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA; [3]Department of Veterinary Pathobiology and [4]Biomedical Genomics Center, University of Minnesota, St. Paul, Minnesota 55108, USA; [5]Human Genome Sciences, Rockville, Maryland 20850, USA; [6]Sanaria Inc., Rockville, Maryland 20852, USA

The apicomplexans *Plasmodium* and *Cryptosporidium* have developed distinctive adaptations via lineage-specific gene loss and gene innovation in the process of diverging from a common parasitic ancestor. The two lineages have acquired distinct but overlapping sets of surface protein adhesion domains typical of animal proteins, but in no case do they share multidomain architectures identical to animals. *Cryptosporidium*, but not *Plasmodium*, possesses an animal-type *O*-linked glycosylation pathway, along with >30 predicted surface proteins having mucin-like segments. The two parasites have notable qualitative differences in conserved protein architectures associated with chromatin dynamics and transcription. *Cryptosporidium* shows considerable reduction in the number of introns and a concomitant loss of spliceosomal machinery components. We also describe additional molecular characteristics distinguishing Apicomplexa from other eukaryotes for which complete genome sequences are available.

[Supplemental material is available online at www.genome.org.]

The availability of two apicomplexan complete genome sequences, *Plasmodium* (Gardner et al. 2002) and *Cryptosporidium* (Abrahamsen et al. 2004), provides a unique opportunity to understand the genome-scale trends accompanying adaptation to parasitic niches in the eukaryotes. All members of the apicomplexan clade are parasitic and share specific features related to parasitism, most notably a unique apical secretory structure mediating locomotion and cellular invasion. Despite general shared features, the apicomplexans have greatly diverged in many respects, including host specificities, tissue tropisms, and the requirement of multiple hosts. Hemosporidians, such as *Plasmodium* and the piroplasms *Theileria* and *Babesia*, infect blood cells and are transmitted to vertebrates by hematophagous arthropod definitive hosts. Most hemosporidians show multiple tissue trophisms and transformation through multiple developmental stages, such as the hepatocyte invasion and intrahepatocytic schizogony that is observed in *Plasmodium*. In contrast, *Cryptosporidium* and the gregarines have a relatively simple parasitic strategy involving a single host and invasion of a single cell type, primarily intestinal epithelial cells.

The current phylogenetic analysis of the characterized Apicomplexa suggest a basal position for *Cryptosporidium* and the gregarines with respect to a poorly defined "crown group" composed of hemosporidians and coccidians (Carreno et al. 1999; Zhu et al. 2000). Hence, comparative genome analysis is likely to yield a picture of both ancestral adaptations common to the characterized Apicomplexa, and also reveal the extent of diversification accompanying the two distant branches of the clade. Furthermore, comparisons with other eukaryotes will provide insights into the affinities of the apicomplexans, and the mode and relative tempo of the evolution of key eukaryotic cellular components. Toward these goals, we present here a comparative analysis of *Plasmodium* and *Cryptosporidium parvum*, with emphasis on comparison of these apicomplexans with eukaryotic lineages with complete genome sequences.

[7]These authors contributed equally to this work.
[8]Corresponding authors.
E-MAIL aravind@ncbi.nlm.nih.gov; FAX (301) 435-7794.
E-MAIL tjt2001@med.cornell.edu; FAX (212) 746-4028.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2615304.

RESULTS AND DISCUSSION

The Relationship of Apicomplexans to Other Eukaryotes and the Degree of Relatedness of the Apicomplexan Proteomes

To obtain a robust phylogenetic model for the relationship of apicomplexans with other eukaryotes having complete genome sequences, we prepared a concatenated multiple alignment (see Supplemental data 1) of >30 conserved *C. parvum* proteins, such as ribosomal proteins, DNA and RNA polymerases, translation factors, and tRNA synthetases having orthologs in *Plasmodium* (Apicomplexa); *Arabidopsis* (plants); *Caenorhabditis*, *Drosophila* and *Homo* (animals); *Neurospora*, *Saccharomyces*, and *Schizosaccharomyces* (fungi); *Giardia* (Parabasalids); and *Aeropyrum* and *Archaeoglobus* (Archaea). This multiple alignment, spanning >4000 aligned positions, was used to compute maximum likelihood, maximum parsimony, neighbor joining, and least squares trees, all rooted using the archaeal sequences. These methods uniformly yield a tree topology with *Plasmodium* and *Cryptosporidium* forming a monophyletic lineage lying outside of a strongly supported "crown group" composed of animals, fungi, and plants (Fig. 1A). *Giardia* occupies a basal position amidst the eukaryotes included in this analysis. This topology is also sup-

14:1686–1695 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

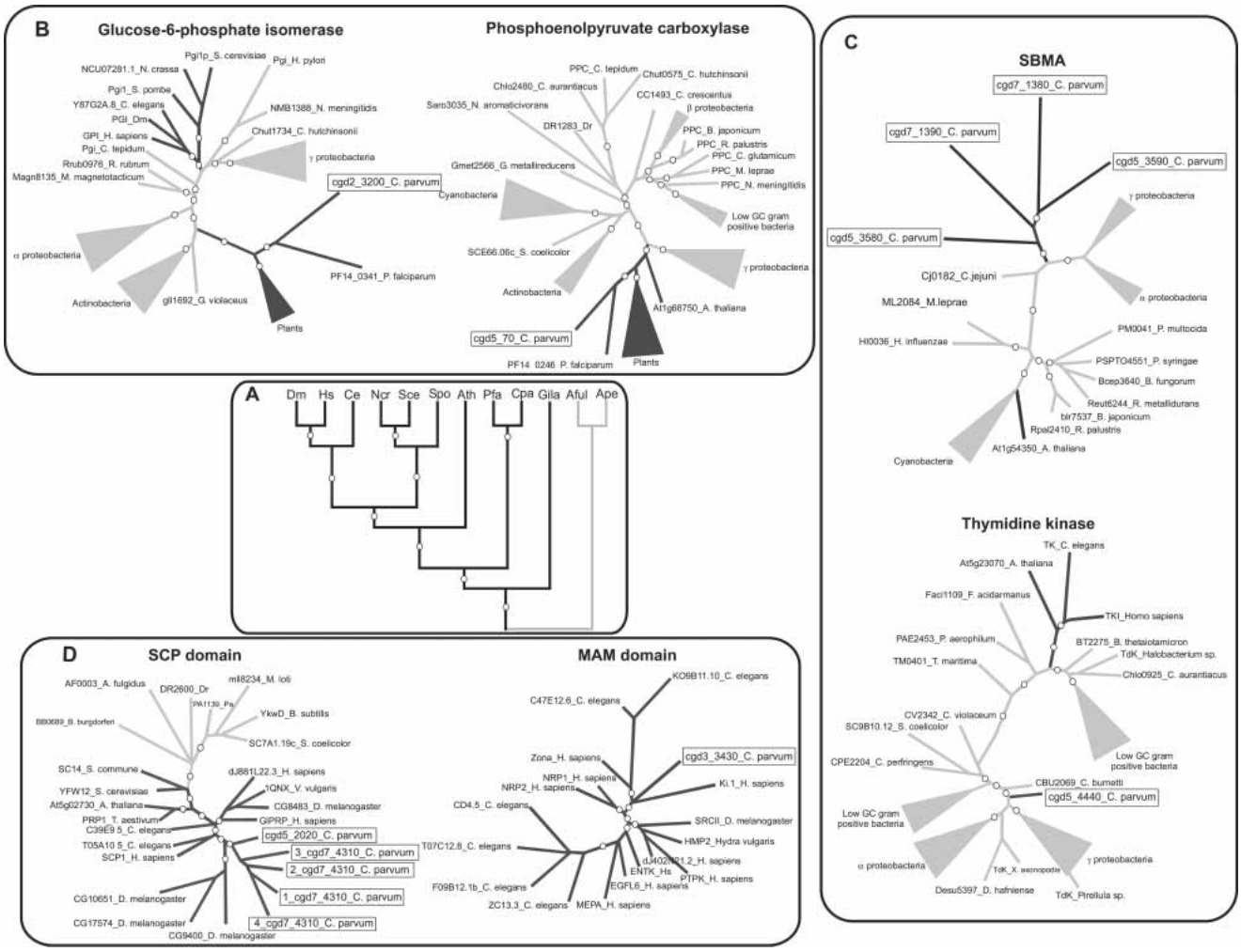

Figure 1 (*A*) Higher-order relationships between eukaryotes (having complete genome sequence information) rooted with archaeal orthologs, as inferred from a concatenated alignment of 30 highly conserved proteins. The circles indicate bootstrap supports >85% (or Bayesian posterior probability > 0.9) obtained by the full ML (PromI), Puzzle ML, weighted neighbor-joining, parsimony, and minimum evolution methods. Bacterial and archaeal branches are in gray and eukaryotic branches are in black. (*B*) Plant affinities of apicomplexan proteins, glucose-6-phosphate isomerase, and phosphoenolpyruvate carboxylase. (*C*) Bacterial affinities of apicomplexan proteins, SBMA and thymidine kinase. (*D*) Animal affinities of apicomplexan proteins, SCP and MAM-domain-containing proteins. In these cases, the circles indicate boostrap support >85% by ML distance analysis (with Puzzle), RellBP, and neighbor-joining methods. Proteins are represented by their gene names and specific names. Some are abbreviated for convenience. Species abbreviations are: (Afu) *Archaeoglobus fulgidus*; (Ape) *Aeropyrum pernix*; (Ath) *Arabidopsis thaliana*; (Bb) *Borrelia burgdorferi*; (Ce) *Caenorhabditis elegans*; (Cpa) *Cryptosporidium parvum*; (Dm) *Drosophila melanogaster*; (Dr) *Deinococcus radiodurans*; (Gila) *Giardia lambia*; (Hs) *Homo sapiens*; (Pa) *Pseudomonas aeruginosa*; (Pfa) *Plasmodium falciparum*; (Sce) *Saccharomyces cerevisiae*; (Spo) *Schizosaccharomyces pombe*.

ported by domain architecture analysis of ~400 proteins belonging to different functional categories as discrete characters. Previous reports propose a weak association between the "plant clade" (comprised of green plants, rhodophytes, and glaucocystophytes) and a large assemblage of eukaryotes that include Stramenopiles and Alveolates (including Apicomplexa; Baldauf et al. 2000). We did not find significant support for a grouping between the alveolates and the plants with the above data set. This suggests that, against the vertical relationship of these eukaryotic taxa, apicomplexan proteins showing specific affinities to plants are likely contributions of the rhodophyte apicoplast progenitor (Fig. 1B). Additionally, several proteins with clear phylogenetic affinity to bacterial homologs were observed (Fig. 1C). These proteins appear to be derived from several distinct bacterial lineages, but the sources of lateral transfers were not always assignable.

To determine the relatedness of the apicomplexan proteome, and to provide a quantitative measurement of proteome similarity and divergence across protein functional categories, we used a simple measure termed the orthology coefficient (OC). For a set of proteins from two compared organisms, the OC represents the fraction occurring in orthologous groups. Thus, an OC = 1 indicates that all the proteins within a compared set have an orthologous relationship. Likewise, if only a fraction of the proteins in the set form orthologous groups, then the OC would fall between 1 and 0. *Plasmodium* and *C. parvum* shared ~2000 orthologous groups, with an overall orthology coefficient of 0.41 (Fig. 2A). This suggests that both parasites possess a significant complement of genes that do not have any orthologous representatives in the other. We further defined OCs for protein sets classified into functional categories (Fig. 2A), revealing that orthology coefficients span a striking range from ~0.2 to 0.4 for proteins related to extracellular adhesion and surface protein glycosylation, and 0.8 to 0.85 for core cellular functions such as translation, RNA processing, ubiquitination, DNA repair/replica-

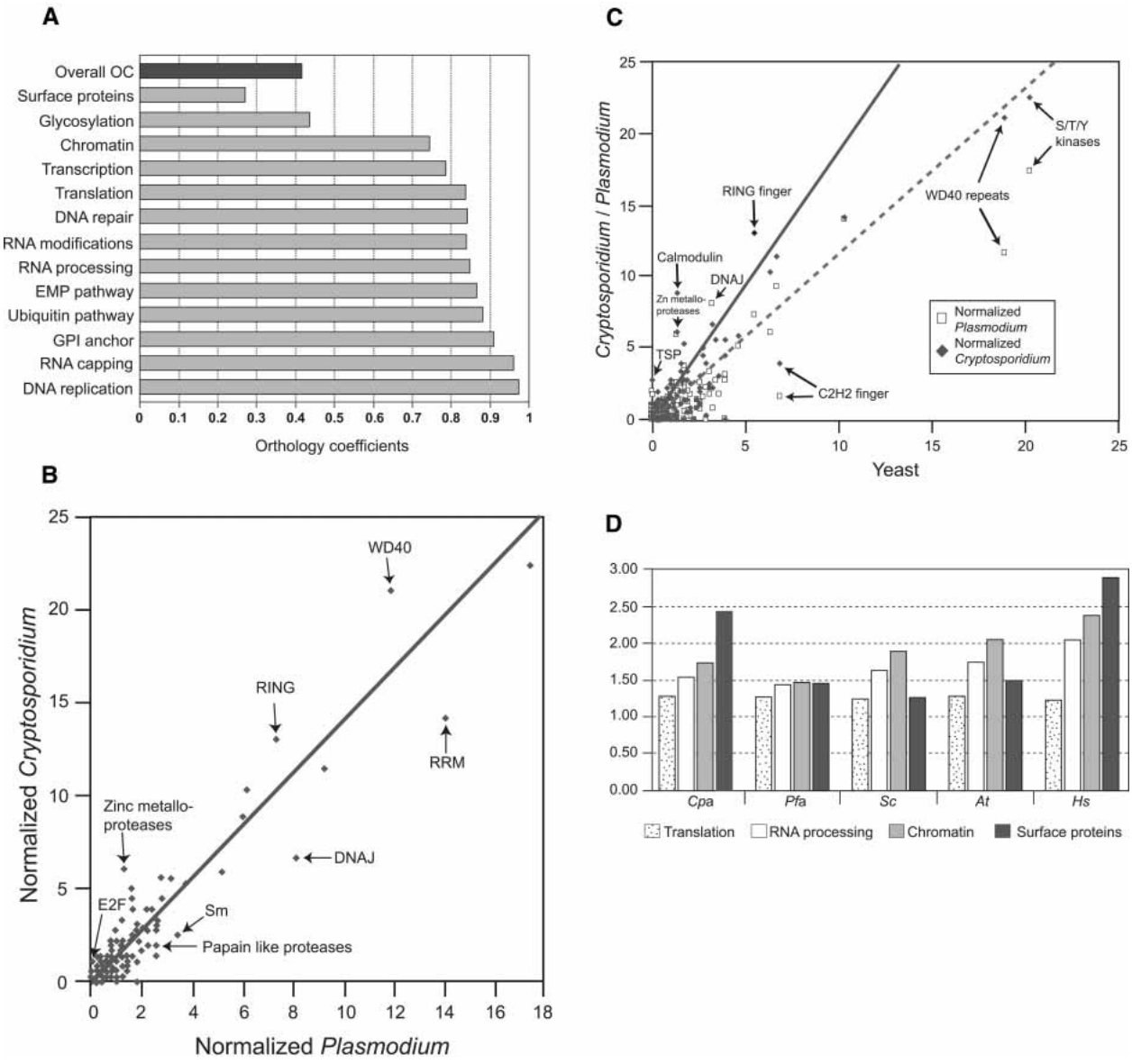

Figure 2 (*A*) Orthology coefficients across different protein functional classes. The overall OC refers to comparison of all proteins within the two apicomplexan proteomes. Surface proteins (or extracellular secreted proteins) are defined as those proteins that contain a predicted signal peptide sequence, lack of ER retention signals, and in many instances, contain transmembrane regions, globular cysteine-rich domains, or known surface protein domains. Note that smaller OC values are observed in the Apicomplexa for functional classes such as chromatin dynamics and splicing. Comparison in different eukaryotes of the number of domains per protein (*B*) and number of types of domains per protein (*C*). (*Cpa*) *Cryptosporidium parvum*; (*Pfa*) *Plasmodium falciparum*; (*Sc*) *Saccharomyces cerevisiae*; (*At*) *Arabidopsis thaliana*; (*Hs*) *Homo sapiens*. Comparison of the demography of the most prevalent conserved domains *Plasmodium* versus *Cryptosporidium* (*D*) and *Cryptosporidium*/*Plasmodium* versus Yeast (*E*) by means of scatterplots. The number of proteins containing an occurrence of 190 commonly found regulatory protein domains were determined in each of the proteome using a library of PSI-BLAST profiles of these domains. The number was then plotted as a scatterplot with each organism being compared representing one of the axes. In each graph, the equivalence lines, which have a slope equal to the ratio of the two proteomes being compared, are shown. Points below the equivalence are overrepresented in the organism on the *x*-axis, whereas points above the equivalence line are overrepresented in the organism on the *y*-axis.

tion, and chromatin dynamics. This suggests that evolutionary divergence of the two parasites has differentially affected various functional classes (Fig. 2A).

Many differences in ortholog distribution could be attributed to gene loss accompanying mitochondrion or apicoplast organellar degradation, or elimination of metabolic pathways in *C. parvum*. For example, versions of the tRNA synthetases and DNA repair enzymes, present in *Plasmodium* but lost in *Cryptosporidium*, likely represent forms with mitochondrion- and apicoplast-specific functions in *Plasmodium*. This prediction, based on patterns of gene loss, is corroborated by the presence of long N-terminal extensions mediating organellar targeting only in the *Plasmodium* versions of these proteins. Qualitative differences between the apicomplexan lineages occur even in the high-OC-value protein sets corresponding to core cellular processes. When the demography of the conserved domains in the two apicomplexan proteomes were compared to each other and against *Saccharomyces cerevisiae*, a unicellular eukaryote with a roughly com-

parable number of protein-coding genes, certain interesting large-scale trends were observed (Fig. 2B,C). The two apicomplexans show independent lineage-specific expansions of entirely different protease families, and *C. parvum* does not share the prominent lineage-specific expansion of RESA-type DnaJ domains that is encountered in *Plasmodium falciparum* (Fig. 2B,C). Relative to yeast, *Cryptosporidium*, like *Plasmodium* (Aravind et al. 2003), also shows a remarkable expansion of the calcium-binding EF hand domains, suggesting that a well-developed calcium-dependent signaling apparatus is likely to have been present in the ancestral apicomplexan. These differences involve both differential gene loss as well as lineage-specific innovations and are discussed further below in the context of specific functional systems. In general, proteins related to functional categories, such as chromatin structure and RNA processing, show fewer domains per protein in the apicomplexan with respect to animals. However, at least in the case of *Cryptosporidium*, the number of domains per protein in modular surface proteins is closer to those in animals (Fig. 2D). A case-by-case examination of these functional categories indicated more pronounced qualitative similarities and differences with protein architectures in other eukaryotes that are potentially correlated with the divergent adaptation of various eukaryotes.

Functional Categories With Low OC Values: Surface Proteins and the Glycosylation Machinery

Previous studies on *P. falciparum* and other apicomplexans have implicated many surface proteins in the recognition of host cells, extracellular matrices, and hemo-lymphatic fluids. Sequence analysis of these molecules has shown that they are distinguishable into two principal classes: (1) those with surface protein domains that are either restricted to a single apicomplexan genus or a few genera of the apicomplexan clade; and (2) those that contain conserved domains widespread across a broad range of organisms in evolution. The former class of proteins includes variant surface antigens (*vars*; Baruch et al. 1995; Peterson et al. 1995; Smith et al. 1995; Su et al. 1995) and the Rifin/Stevor family (Cheng et al. 1998; Gardner et al. 1998) from *Plasmodium*, and oocyst wall proteins (OWPs, also present in *Toxoplasma*; Templeton et al. 2004) and the newly described lineage-specific surface molecules (Abrahamsen et al. 2004) of *Cryptosporidium*. These proteins are characterized by predominant α-helical compositions, or cysteine-rich modules stabilized by disulfide bridges, suggesting that they have emerged rather late in evolution in particular lineages of apicomplexans. Extensive proliferation and divergence of these proteins are likely caused by selective forces such as host immune pressure driving antigenic diversity. Exemplifying the second class of proteins are MSP1, *P*25/28, CSP1, and TRAP from *P. falciparum* and the *Toxoplasma gondii* MIC micronemal proteins. These proteins contain conserved adhesion domains, such as the EGF domain, Thrombospondin type 1 (TSP1) domain, the von Willebrand Factor A (vWA) domain, and the APPLE domain, that are typically abundant in animal surface proteins but are either absent or rarely present in surface adhesion molecules in other eukaryotic lineages examined to date.

We systematically investigated the affinities of the surface protein domains by searching the *C. parvum* proteome with a comprehensive library of PSI-BLAST-derived position-specific score matrices and hidden Markov models for surface protein domains. These profiles were previously used to detect such domains in adhesion proteins of *Caenorhabditis elegans*, *Homo sapiens*, and *P. falciparum* (Aravind and Subramanian 1999; Lander et al. 2001; Chervitz et al. 1998). As a result of these analyses, we have identified 32 widely conserved surface domains distributed in 51 proteins (Supplemental Table 1), including 24 noncatalytic protein- or carbohydrate-interacting domains and seven catalytic domains (see domain architectures in Fig. 3A). Most strikingly, 10 of these domains, namely, TSP1, Sushi/CCP, Notch/Lin1 (NL1), NEC (Neurexin-Collagen domain), Fibronectin type 2 (FN2), Pentraxin, MAM, ephrin receptor EGF-like domain, the animal signaling protein hedgehog-type HINT domain, and the Scavenger receptor domain, have thus far been found only in the surface proteins of animals other than apicomplexans. The remaining domains, such as the EGF, LCCL, Kazal-type protease inhibitor, Kringle, ToxI, PR1/SCP, and Fibronectin type 3 (FN3) domains, are seen in other eukaryotes, but their extracellular forms are predominantly found only in animals (Supplemental Table 1; Fig. 1D). Additional sets of domains from the apicomplexan surface proteins, namely, the clostripain protease domain (a rare divergent protease domain of the caspase-hemoglobinase fold), levanase-associated lectin-, the discoidin-, *Archaeoglobus* protease associated cysteine rich-, and the anthrax toxin subunit N-terminal domains (Supplemental Table 1; Fig. 3A), show clear prokaryotic affinities.

The surface protein adhesion domains in the apicomplexan proteomes can be attributed to multiple distinct heritages: those originally derived from bacteria and animals and laterally transferred to Apicomplexa, and those "invented" within the Apicomplexa, typically in a lineage-specific manner. In principle, it is possible that the surface protein domains shared by animals and apicomplexans were present in the ancestral eukaryotes and secondarily lost in other lineages. Although gene loss occurs frequently in eukaryotes, most of these domains shared by both these lineages are often undetectable in the (nearly) completely sequenced genomes of multicellular (filamentous) fungi, plants, and other unicellular eukaryotes such as trypanosomes and parabasalids. Thus, if multiple gene losses were to be invoked, it would imply that the common ancestor of these lineages was probably more diverse in its protein complement than most of the descendents, and this is not consistent with the demography of the protein families encoded by eukaryotic genomes (Lespinet et al. 2002). Furthermore, in phylogenetic analysis, specific affinities between apicomplexan and animal versions were recovered (Fig. 1D; Pradel et al. 2004). The animal affinities of domains are also highly overrepresented in the category of surface proteins, as against intracellular proteins belonging to other functional categories. Taken together, these observations make lateral transfer from animals, followed by selective retention of functionally relevant protein domains involved in adhesion, as the most parsimonious explanation for these observations.

The plausibility of horizontal transfer from an animal source is also supported by the intracellular location of apicomplexan parasites for most of their life cycle and, in the case of *Plasmodium*, there is facile uptake and expression of DNA constructs introduced to the erythrocyte cytoplasm prior to parasite infection (Deitsch et al. 2001). The bacterial component of the lateral transfer could be attributed, in part, to the original cyanobacterial origin of the apicoplast progenitor. Additionally, it is possible that the apicomplexans acquired additional bacterial genes through contact with bacteria cohabitating their ecological niches, such as animal guts and intracellular niches. Interestingly, the domain architectures of these apicomplexan surface proteins are more similar to those present in the multicellular organisms like animals, in terms of the numbers and diversity of domains, than proteins present in bacteria or unicellular eukaryotes, such as yeasts and microsporidians. This, taken together with the observation that bacterial parasites have acquired far fewer animal surface proteins (Ponting et al. 1999; Subramanian et al. 2000), suggests that the presence of eukaryotic secretory

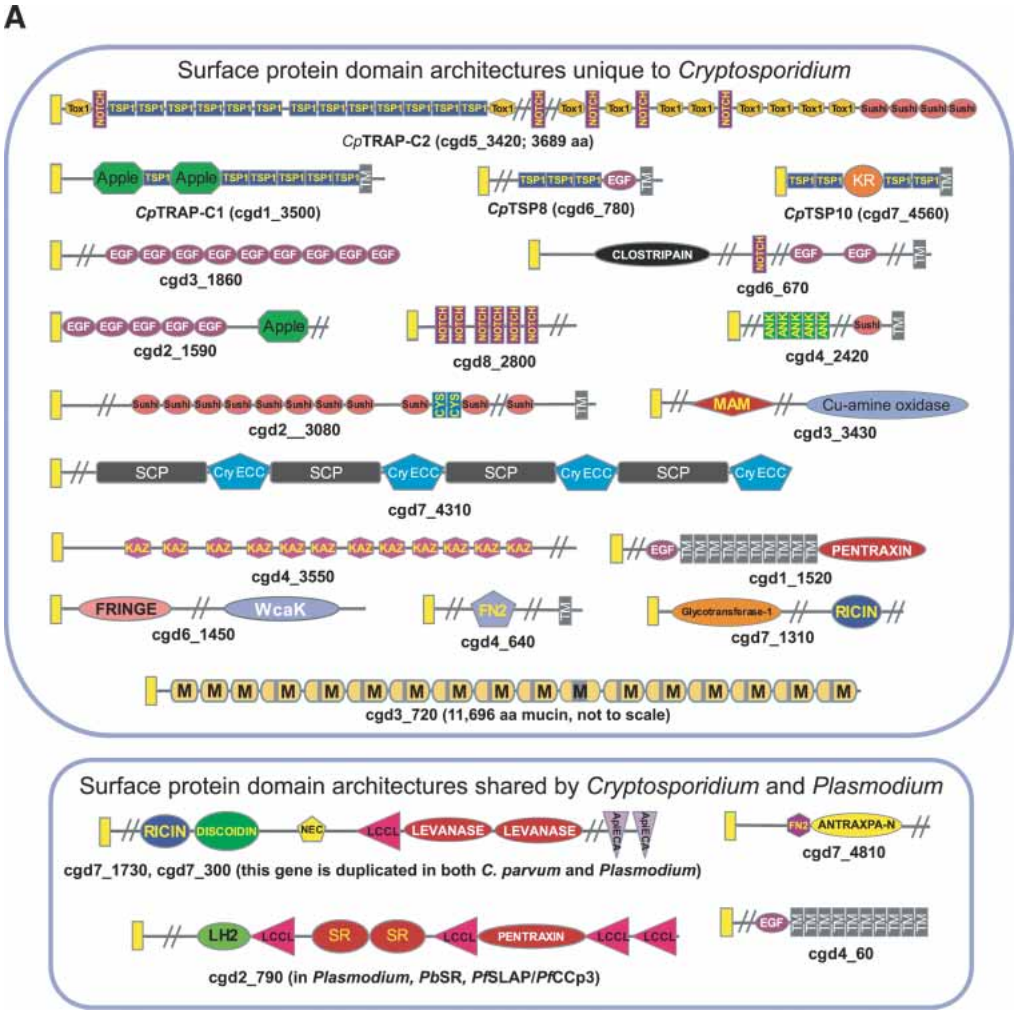

Figure 3 (Continued on next page)

and glycosylation systems in the apicomplexans facilitated utilization of laterally acquired domains of animal provenance.

For *P. falciparum* and *C. parvum* surface proteins, the OC is no more than 0.2 for proteins having conserved "animal-type" and "bacterial-type" domains, with few conserved architectures shared by the two lineages. Furthermore, the set of surface domains in these proteins is overlapping but not identical in the apicomplexans. For example, the MAC/perforin-type domain (Aravind et al. 2003) is found only in *Plasmodium* whereas the animal Hedgehog-related HINT domain (Hall et al. 1997) is found only in *Cryptosporidium*. This suggests that, although some lateral transfer events may have occurred early in the evolution of the apicomplexan clade, extensive lineage-specific domain acquisition, gene loss, and domain shuffling occurred during speciation. In most animals, these surface protein domains are encoded by distinct exons and the architectural diversity arises through exon shuffling. In contrast, the majority of *C. parvum* multidomain proteins are encoded within a single exon, whereas in *Plasmodium* the multiexon genes show no clear correlation with the structure of the domain architecture of the protein. This suggests that the process of domain shuffling in the Apicomplexa is likely unrelated to the exon shuffling process of animal multidomain surface proteins.

Comparison of the surface protein glycosylation apparatus reveals dramatic divergence between these two apicomplexans (Fig. 3B). Both possess a well-developed GPI anchor synthesis apparatus that is largely similar to the corresponding pathway in other eukaryotes. Unlike *Plasmodium*, *Cryptosporidium* lacks the canonical *N*-acetylglucosaminylphosphatidylinositol deacetylase that catalyzes the second step in the GPI anchor biosynthetic pathway. However, sequence analysis revealed the presence of an unrelated bacterial-type sugar deacetylase (cgd1_3060) that is likely to catalyze the same reaction. Whereas there is only a rudimentary *N*-linked glycosylation pathway present in *Plasmodium*, a more developed pathway is predicted in *Cryptosporidium*. *N*-linked glycosylation has been widely detected in other eukaryotes, including *Toxoplasma* (Odenthal-Schnittler et al. 1993), and it is likely that *Cryptosporidium* retains the primitive state whereas most of the apparatus has degenerated in *Plasmodium*. In stark contrast to *Plasmodium*, we detected at least seven enzymes of the canonical *O*-linked glycosylation pathway in *C. parvum* (Fig. 3B). The core enzymes for this pathway have previously only been observed in the animals (Varki 1999). It is therefore possible that the *Cryptosporidium* lineage acquired this capacity from an animal host at some point during its evolution. Interestingly, one of the galactosyl transferases of this pathway, which is homologous to the animal Fringe protein, contains an additional C-terminal domain related to the bacterial WcaK-like glycosyltransferase domains (Reeves et al. 1996). *Cryptosporidium* also possesses a second standalone version of the WcaK-like glycosyltransferase,

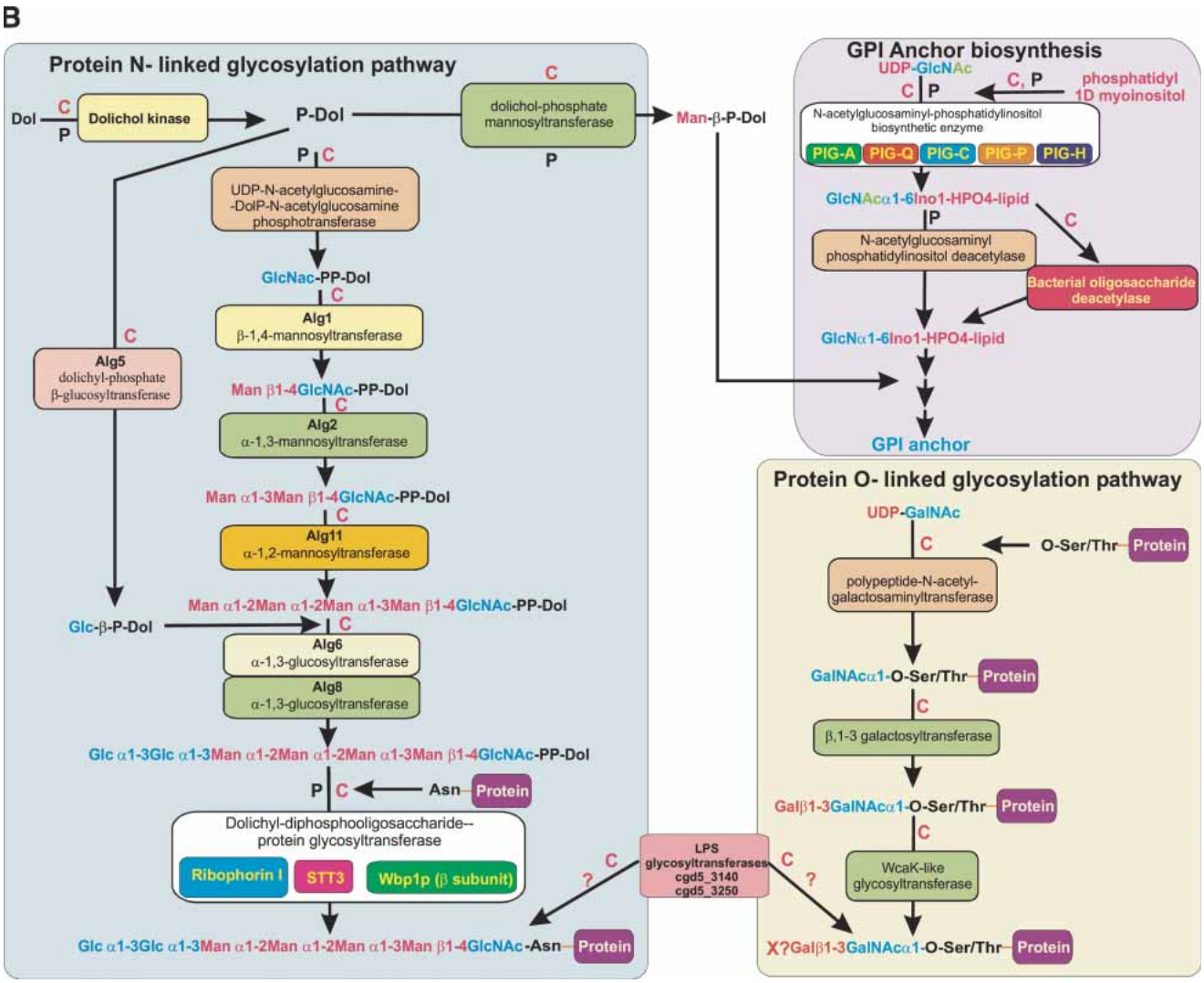

Figure 3 (*A*) Domain organizations of a representative set of surface proteins from *Cryptosporidium parvum* (*top* panel) and orthologs common to *Plasmodium falciparum* (*bottom* panel). All proteins shown here have a signal peptide sequence represented by a yellow rectangular box at the beginning of the architectures. The domains are labeled as in Supplemental Table 1. Those not shown in Supplemental Table 1 are (Ank) ankyrin repeat; (CYS) cysteine-rich repeats found in *Archaeoglobus* proteases; (M) mucophorin domain (with Thr/Ser stretches indicated by gray boxes; see Supplemental Fig. 1); and (TM) membrane-spanning region. (*B*) Schematic representation of the reconstructed glycosylation pathways in Apicomplexa. The enzymes are shown in boxes along with the protein names of the respective yeast homologs. The reconstructed oligosaccharide chain is shown using abbreviations for the various sugars. (Glc) Glucose; (Gal) galactose; (Man) mannose; (GlcNAC) *N*-acetylglucosamine; (GalNAC) *N*-acetylgalactosamine; (Dol) dolichol; and (Ino) inositol. (X?) The uncharacterized sugar added by the WcaK-like glycosyltransferases. Wherever *Cryptosporidium* contains an enzyme of the pathway, it is indicated with a C in red, and *Plasmodium* is indicated with a P in black.

which may decorate the core *O*-linked oligosaccharides with sugar moieties unique to the parasite. Consistent with the presence of an *O*-linked glycosylation pathway, we detected >30 mucin-like surface proteins in *C. parvum* having stretches of serines and threonines in their extracellular regions, possibly functioning to mediate adhesive interactions with the host cell surface. Remarkable among the mucins is an 11,696-amino-acid protein (cgd3_720; gi: 46228293) having an architecture largely composed of 17 repeats of an ~600-residue-long *C. parvum*-specific all-β-strand globular 12-cysteine domain (Fig. 3A; an alignment of the repeated domain is shown in Supplemental Fig. 1). The 14 C-terminal-most domains each have a predicted internal loop containing a Thr/Ser stretch that is likely a target for glycosylation, including one domain containing 360 consecutive Ser/Thr residues (domain 11, indicated in Fig. 3A and Supplemental Fig. 1).

Low OC of Metabolic Pathway Components Suggest Life-Cycle-Specific Adaptations

The fairly low OC value for the metabolic machinery is understandable given that the *C. parvum* metabolism is greatly streamlined in comparison to *Plasmodium* (Abrahamsen et al. 2004). In noted contrast, *C. parvum* possesses specialized pathways absent in *Plasmodium*, such as the presence of at least nine enzymes related to the metabolism of high-molecular-weight polysaccharides, glycogen or amylopectin. These include biosynthetic enzymes such as glycogen phosphorylase, storage proteins such as amylopectin/starch-binding proteins, and catabolic enzymes including amylases and debranching enzymes. Interestingly, we also detected an ortholog of the plant starch-associated protein R1, an α-glucan, water dikinase (Ritte et al. 2002). This enzyme has otherwise been detected only in the plant lineage, and was

therefore probably acquired from the genome of the rhodophyte apicoplast progenitor. The presence of glycogen/amylopectin in a range of protists, including ciliates and dinoflagellates (Laybourn-Parry 1984), suggests that polysaccharide synthesis is an ancestral adaptation related to food-storage accompanying cyst formation. Loss of this primitive polysaccharide metabolism pathway in the hemosporidians may have occurred following the emergence of an insect vector, along with the elimination of external cyst stages.

The stem of the glycolytic pathway represents the most highly conserved metabolic pathway between *Cryptosporidium* and *Plasmodium*. However, unlike *Plasmodium*, *Cryptosporidium* also possesses enzymes for the terminal metabolism of pyruvate such as pyruvate:ferredoxin oxidoreductase, pyruvate decarboxylase, and malate dehydrogenase. Phylogenetic analysis of pathway components reveals a mosaic of strong affinities to enzyme versions of plants and bacteria. For example, the apicomplexan phosphoglucomutase, phosphofructokinase, and enolase enzymes grouped with the plant versions, whereas fructose bisphosphate phosphatase and phosphoglucomutase showed bacterial affinities. These affinities suggest displacements of the ancestral eukaryotic enzymes by versions derived from the apicoplast precursor and bacterial sources. Nevertheless, the current state of the data precludes us from determining the temporal point in alveolate evolution at which these displacements occurred. Interestingly, similar to the parasite *Leishmania*, *C. parvum* possesses a plant-type 2-phosphoglycerate kinase implicated in archaea in the synthesis of the possible denaturation protectant, 2–3 cyclic phosphoglyceric acid (Matussek et al. 1998).

Differences Between *Cryptosporidium* and *Plasmodium* in Functional Classes With Moderate to High OCs, and Comparisons With Other Eukaryotes

Despite having moderate to high orthology coefficients, functional classes such as RNA processing/splicing and chromatin dynamics provide a picture of how even well-conserved functions can be affected by the divergence of two lineages from a common ancestor. Striking numerical difference is seen in the complements of two RNA-binding domains, Sm and RRM, between *P. falciparum* (17 and 71 domains, respectively) and *C. parvum* (9 and 51 domains; Fig. 2B,C). In particular, *C. parvum* has lost genes encoding Sm domain proteins associated with the U4/U6 and U4/U6 · U5 snRNPs spliceosomal particle, suggesting that the particle activity has degenerated in this organism. The reduction in the number of RRMs also results, in part, from the loss of conserved proteins belonging to the spliceosomal machinery (Fig. 4). Consistent with this loss, the number of predicted introns in *Cryptosporidium* (<10% of genes harbor introns) is vastly lower than those seen in *Plasmodium* (>50% of genes harbor introns; Gardner et al. 2002). This situation is reminiscent of the similar degeneration of the splicing machinery in *S. cerevisiae* versus *Schizosaccharomyces pombe* (Aravind et al. 2000), suggesting that on multiple occasions in eukaryotic evolution, the loss of introns has triggered degeneration of the splicing machinery.

The ratio of the total number of proteins in the proteome to the predicted specific transcription factors in *Cryptosporidium* and *Plasmodium*, 340 and 800, respectively, is in great contrast to the ratio of 29 in *S. cerevisiae*. The decreased ratio in *C. parvum* relative to *Plasmodium* is caused by a greater absolute number of specific transcription factors possessing a variety of conserved DNA-binding domains, such as E2F/DP1, bZip, and GATA DNA-binding domains, in conjunction with a lower overall gene count (Fig. 2B,C). Nevertheless, the numbers of specific transcription factors are far fewer than those encountered in yeast and other eukaryotes, suggesting major differences in the mechanisms of apicomplexan gene regulation. Recent microarray studies on *P. falciparum* indicate a continuous cascade of gene expression in the course of its intraerythrocytic stage cycle, in which groups of functionally related genes are coexpressed, with those involved in generalized functions being expressed first followed by those for increasingly specialized, lineage-specific functions (Bozdech et al. 2003; Le Roch et al. 2003). The relative paucity of transcription factors in both the apicomplexans suggests that the regulation of such transcriptional cascades may be dependent on the well-developed chromatin-remodeling apparatus (Fig. 4), which might work in conjunction with the small set of specific DNA-binding proteins detected in these genomes. In this context, it is of interest to note that both of these apicomplexans contain orthologs of the DNA cytosine methyltransferase DNMT2 (the *P. falciparum*, gi: 23612639, and *C. parvum* orthologs were recovered with *e*-values of 10^{-15} and 10^{-26}, respectively, in searches of the nonredundant protein database). These apicomplexan DNMT2 orthologs showed a full-length alignment with the versions from other eukaryotes (see alignment, Supplemental Fig. 2), and the sequence similarity encompassed both the N-terminal catalytic Ado-Met binding domain and a C-terminal domain unique to the DNMT2 family of methylases (Tang et al. 2003). They also show absolute conservation of the active-site residues implicated in Ado-Met binding and catalysis, suggesting that they are active enzymes. Consistent with this, cytosine methylation has been previously reported in *Plasmodium* (Pollack et al. 1991); however, it remains to be determined if they participate in an epigenetic control mechanism related to transcriptional cascading.

Although there is a much higher correspondence between the two apicomplexans in chromatin proteins than specific transcription factors, interesting differences are found in both absolute numbers and architectures of these proteins (Fig. 4). *C. parvum* has 14 chromatin-remodeling SNF2/SWI2 ATPases, whereas *Plasmodium* has just 11. Comparisons of the apicomplexan Swi2/Snf2 ATPases with the other eukaryotes suggests that *Plasmodium* has lost the Rad26 and Swr1 orthologs, whereas *Cryptosporidium* appears to have lost one of the Rad16/Rad5-like Ring finger-containing forms and the version fused to a C-terminal Endonuclease VII domain. *Cryptosporidium* possesses a version of the Swi2/Snf2 ATPase, with a unique architecture containing two N-terminal chromo domains and one bromo domain. Likewise, *Plasmodium* possesses a unique predicted chromatin-associated protein, having an amine oxidase domain fused to a C-terminal PHD finger that is predicted to function as a novel enzyme that might modify histone amino groups or chromatin basic amines (Fig. 4). It may represent a case of an independently derived chromatin-associated oxidase, parallel to the amine oxidase fused to the SWIRM domain that is seen in the crown group eukaryotes (Aravind and Iyer 2002). The two apicomplexans also share several chromatin proteins with domain architectures unique from any of the crown group eukaryotes. Examples of these include an ISWI-related Swi2/Snf2 ATPase with five PHD fingers, a protein that combines bromo domains with N-terminal ankyrin repeats, and four distinct SET domain methylases joined to a variety of other protein- and nucleic-acid-interacting domains (Fig. 4). An exploration of apicomplexan nuclear proteins having novel architectures may shine light on epigenetic regulatory mechanisms unique to this lineage.

The pellicle in several protozoans is supported by a distinctive fibrous cytoskeletal structure predominantly composed of low complexity, proline- and valine-rich proteins called articulins (Mann and Beckers 2001). We identified 10 distinct articulins in *Plasmodium* and six in *Cryptosporidium* having composition and sequence similarity to the articulins from other alveolates such as the ciliate, *Pseudomicrothorax*. This suggests the

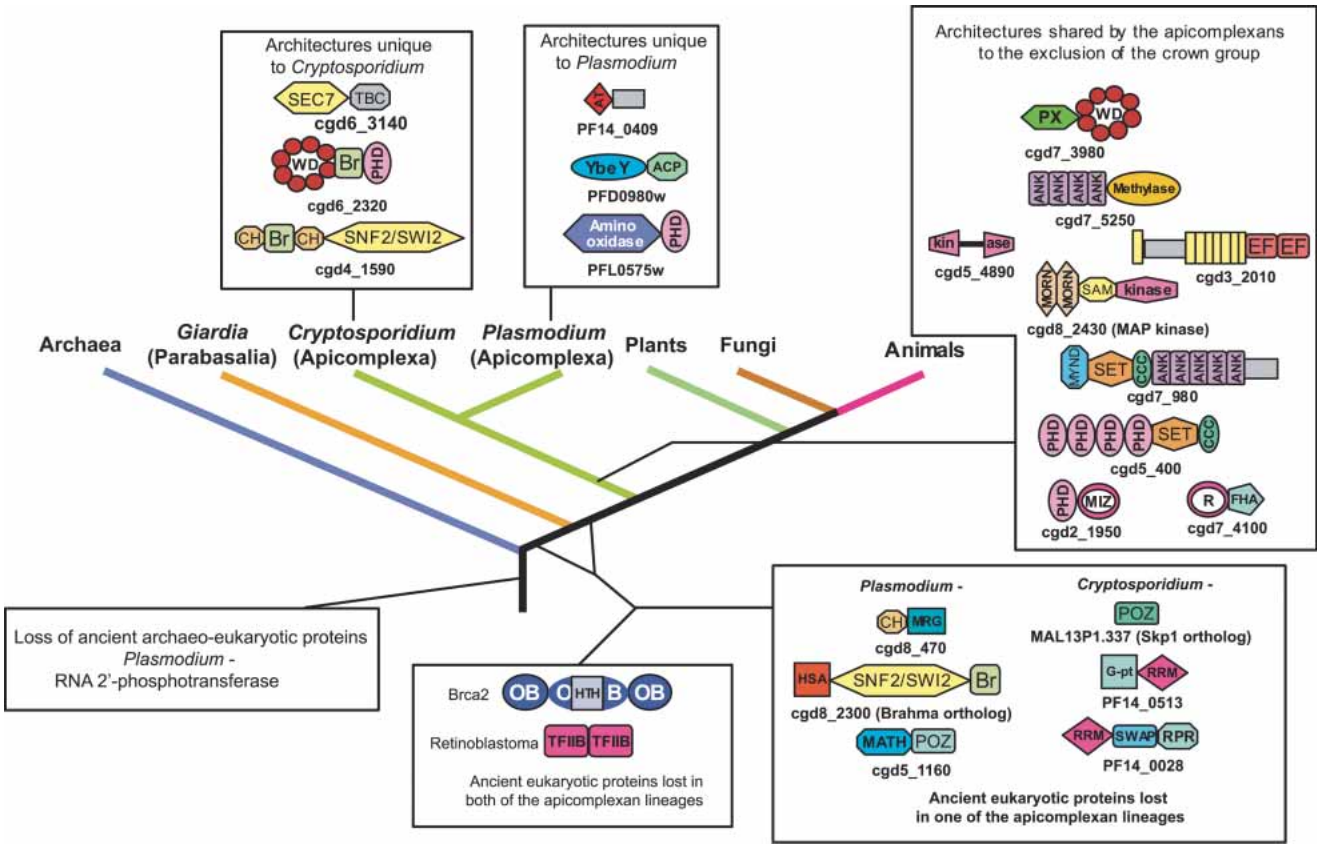

Figure 4 Eukaryotic tree showing select points of derivation and loss of various architectures. The proteins are designated by either *Plasmodium* or *Cryptosporidium* gene names shown below a cartoon of their architecture. Gray boxes indicate globular domains that are not detected elsewhere. Yellow boxes indicate transmembrane segments or signal peptides. The domains found in chromatin proteins are (Ch) chromo domain; (Br) bromo domain; (PHD) PHD finger; (SET) SET protein methyltransferase domain; (CCC) cysteine cluster associated with SET domains; (AT) AT hook domain; (HTH) helix–turn–helix domain; (MYB) MYB-type HTH domain; (TFIIB) TFIIB-like HTH domain; (OB) oligomer-binding domain; (SWI2/SNF2) ATPase module of chromatin-remodeling proteins; (HAS) domain found in SWI2/SNF2 ATPases. The signaling domains are MYND and MIZ-Zn-finger domains; (R) ring finger domain; (ANK) ankyrin domain; (WD) WD40 β-propeller domain; (Kinase) protein kinase domain; (SAM) sterile α-motif domain; (Sec7) ARF GTPase exchange factor domain; (TBC) GTPase-activating domain; (MORN) a β-hairpin repeat motif; (POZ) pox virus zinc finger domain; (MATH) meprin-A5-TRAF homology domain; (EF) EF-hand domain. RNA-binding domains are (G-patch) glycine-containing RNA-binding domain; (SWAP) suppressor of white apricot domain; (RRM) RNA recognition motif domain. Other domains are (YbeY) predicted metal-dependent lecithinase domain; (ACP) acyl carrier protein domain. cgd8_2430 is a predicted MAP kinase, cgd5_4390 kinase shows a lineage-specific expansion in *Plasmodium*, and cgd3_2010 is predicted to be a novel signaling receptor with intracellular calcium-binding EF-hand domains.

retention of this ancient cytoskeletal feature of the alveolate clade despite the dramatic parasitic adaptations of the apicomplexans. Our current analysis also identifies as an articulin the *Plasmodium* gametocyte-expressed protein, *Pf*s77 (Baker et al. 1995), suggesting involvement of articulins in the maintenance of stage-specific cellular shapes.

Conclusions

Comparison of the *Plasmodium* and *Cryptosporidium* complete genome sequences reveals that the ancestral apicomplexan encoded at least 145 shared "apicomplexan" proteins with no obvious orthologs in other organisms (see Supplemental data 2). This apicomplexan set includes ~30 membrane proteins and five secreted proteins. These surface proteins, unique to the apicomplexan lineage, possibly participate in the formation of surface structures related to interactions with eukaryotic host cells and the biogenesis of the apical complex. The unique intracellular apicomplexan proteins, which are typically enriched in low-complexity segments, are also likely to be internal structural components of lineage-specific organelles such as dense granules, micronemes, rhoptries, and the apical complex. In contrast to other eukaryotic parasites such as the microsporidians, kinetoplastids, and *Giardia*, the apicomplexans show a wide array of surface proteins with domains that are typically prevalent in animal surface proteins. At least five multidomain proteins (Fig. 3A, lower panel) can be traced back to the common ancestor of the two apicomplexan lineages, suggesting that the ancestor had already acquired a set of domains from a host belonging to the animal lineage.

However, beyond the core set of genes, the evolution of parasitism has involved lineage-specific adaptation occurring through gene loss, additional lateral transfers, and lineage-specific expansions. Considerable lineage-specific gene loss is indicated by the absence of widespread eukaryotic proteins specifically in either one of the apicomplexan lineages (Fig. 4), suggesting that the common ancestral apicomplexan possessed a more complex genome encoding a greater repertoire of biochemical activities. The streamlining appears to correlate with increasing propensity of the parasite to use its host(s) for most of its metabolic requirements. Some of the losses are not easily explained: for example, in *Cryptosporidium*, the apparent elimination of Skp1p and Skp2p-like proteins, despite the presence of cullins, suggests potential differences in the cell cycle related ubiquitination complexes relative to other eukaryotes. Thus, as a conse-

quence of gene losses and lineage-specific innovations, apicomplexans possess proteomes quite different from free-living unicellular eukaryotes that have similar overall gene numbers (Fig. 2A–D). Most notably, the apicomplexan proteomes have a large component devoted to pathogenesis, immune evasion, and adhesion rather than transcription, posttranscriptional regulation, or metabolism.

The apicomplexans confirm large-scale trends in the evolution of the eukaryotes, specifically the involvement of lineage-specific expansions in the generation of specific adaptations (Abrahamsen et al. 2004) and lineage-specific architectural diversification of proteins and functions using conserved pools of domains, such as those involved in signal transduction, chromatin dynamics, and transcription. Experimental investigation of the architectural variations may enlighten fundamental aspects of biological diversification.

METHODS

Sequence Analysis and Phylogenetic Tree Constructions

C. parvum genome sequence information and annotation supporting this manuscript are available on an in-house genome browser (http://134.84.110.219/cgi-bin/gbrowse/crypto909) and at (http://www.cryptodb.org). General methodologies supporting genome sequence annotation, including BLAST searches, multiple sequence alignments, protein structure determinations, gene family clustering, and phylogenetic analyses were performed as briefly follows. The nonredundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, November 21, 2003) was searched using the BLASTP program. Profile searches were conducted using the PSI-BLAST program (Altschul et al. 1997) with either a single sequence or an alignment used as the query, with a default profile inclusion expectation (*E*) value threshold of 0.01 (unless specified otherwise), and was iterated until convergence. Multiple alignments were constructed using the T_Coffee program (Notredame et al. 2000), followed by manual correction based on the PSI-BLAST results. Signal peptides were predicted using the SIGNALP program (http://www.cbs.dtu.dk/services/SignalP-2.0/; Nielsen et al. 1997). Transmembrane regions were predicted in individual proteins using the PHDhtm (Rost et al. 1996), TMHMM2.0 (Krogh et al. 2001), and TOPRED1.0 (Claros and von Heijne 1994) programs with default parameters. For TOPRED1.0, the organism parameter was set to "eukaryote" (http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html). Additionally, the multiple alignments were used to predict TM regions with the PHDhtm program. The library of profiles for conserved protein domains were also prepared by extracting alignments from the PFAM database (Bateman et al. 2002; http://www.sanger.ac.uk/Software/Pfam/index.shtml) and updated by adding new members from the NR database. These updated alignments were then used to make HMMs with the HMMER package (Eddy 1998; Bateman et al. 2002) or PSSM with PSI-BLAST. All large-scale sequence analysis procedures were carried out using the SEALS package (http://www.ncbi.nlm.nih.gov/CBBresearch/Walker/SEALS/index.html). Similarity-based clustering of proteins was carried out using the BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.txt).

Phylogenetic analyses were carried out using the maximum-likelihood, neighbor-joining, protein parsimony, and least-squares methods (Felsenstein 1989, 1996; Hasegawa et al. 1991). The protein parsimony was carried out only for the concatenated alignment of conserved eukaryotic and archaeal proteins. Parsimony and neighbor-joining analysis were carried out using the Mega package (Kumar et al. 2001). Additionally, weighted neighbor-joining trees with corrections for long branch effects using the WEIGHBOR program (Bruno et al. 2000). When the total number of taxa was manageable (including the concatenated alignment), we constructed full maximum likelihood (ML) trees using the Proml program (100 bootstrap replicates and global rearrangements) of the Phylip package. TreePuzzle 4.02 (Schmidt et al. 2002) was used to estimate the parameters with a γ correction for among site rate variation plus a correction for invariant sites (8 + 1 rate categories) from the data sets. ML distance analyses used TreePuzzle 4.02 to calculate ML distance matrices along with Puzzleboot for 1000 replicates (http://www.tree-puzzle.de); resampled matrices were then analyzed using Fitch (with global rearrangements and 10 times jumbling) from the Phylip package, and the WEIGHBOR program. In an alternative method, a least-squares tree was constructed with the FITCH program (from the Phylip package; Felsenstein 1989) followed by local rearrangement using the Protml program of the Molphy package (Hasegawa et al. 1991) to arrive at the maximum likelihood (ML) tree. The statistical significance of various nodes of this ML tree was assessed using the relative estimate of logarithmic likelihood bootstrap (Protml RELL-BP), with 10,000 replicates. The Bayesian posterior probability trees were also constructed for the data set using the MrBayes 3 program (Ronquist and Huelsenbeck 2003). The test for alternative phylogenetic hypothesis was performed using the Consel program.

Calculation of Orthology Coefficients (OC)

Orthology coefficients were calculated as described (Subramanian et al. 2000) and as follows. In the case of a one-to-one correspondence between genes in two genomes, OC = 2No/(N1 + N2), where No is the number of orthologs and N1 and N2 are the numbers of members of the given protein family or functional category in the two compared genomes. If there is a duplication (two or more members) in one or both of the species that occurred after divergence of the two species, then OC = (No1 + No2)/(N1 + N2), where No1 and No2 are the numbers of members in orthologous clusters from the two respective genomes (Subramanian et al. 2000).

ACKNOWLEDGMENTS

This work was supported in part by the Niarchos Foundation. The Department of Microbiology and Immunology at Weill Medical College acknowledges the support of the William Randolph Hearst Foundation. This study utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, MD. (http://biowulf.nih.gov).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A., Deng, M., Liu, C., Widmer, G., Tzipori, Z., et al. 2004. The complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304:** 441–445.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Aravind, L. and Iyer, L.M. 2002. The SWIRM domain: A conserved module found in chromosomal proteins points to novel chromatin-modifying activities. *Genome Biol.* **3:** RESEARCH0039.

Aravind, L. and Subramanian, G. 1999. Origin of multicellular eukaryotes—Insights from proteome comparisons. *Curr. Opin. Genet. Dev.* **9:** 688–694.

Aravind, L., Watanabe, H., Lipman, D.J., and Koonin, E.V. 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci.* **97:** 11319–11324.

Aravind, L., Iyer, L.M., Wellems, T.E., and Miller, L.H. 2003. *Plasmodium* biology: Genomic gleanings. *Cell* **115:** 771–785.

Baker, D.A., Thompson, J., Daramola, O.O., Carlton, J.M., and Targett, G.A. 1995. Sexual-stage-specific RNA expression of a new *Plasmodium falciparum* gene detected by in situ hybridization. *Mol. Biochem. Parasitol.* **72:** 193–201.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290:** 972–977.

Baruch, D.I., Pasloske, B.L., Singh, H.B., Bi, X., Ma, X.C., Feldman, M., Taraschi, T.F., and Howard, R.J. 1995. Cloning the *P. falciparum* gene encoding *Pf*EMP1, a malarial antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82:** 77–87.
Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res*. **30:** 2762–2800.
Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L. 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol*. **1:** E5.
Bruno, W.J., Socci, N.D., and Halpern, A.L. 2000. Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol*. **17:** 189–197.
Carreno, R.A., Martin, D.S., and Barta, J.R. 1999. *Cryptosporidium* is more closely related to the gregarines than to coccidia as shown by phylogenetic analysis of Apicomplexan parasites inferred using small-subunit ribosomal RNA gene sequences. *Parasitol. Res*. **85:** 899–904.
Cheng, Q., Cloonan, N., Fischer, K., Thompson, J., Waine, G., Lanzer, M., and Saul, A. 1998. *stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. *Mol. Biochem. Parasitol*. **97:** 161–176.
Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S.S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., et al. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282:** 2022–2028.
Claros, M.G. and von Heijne, G. 1994. TopPred II: An improved software for membrane protein structure predictions. *Comput. Appl. Biosci*. **10:** 685–686.
Deitsch, K., Driskill, C., and Wellems, T. 2001. Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes. *Nucleic Acids Res*. **29:** 850–853.
Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.
Felsenstein, J. 1989. PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics* **5:** 164–166.
———. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol*. **266:** 418–427.
Gardner, M.J., Tettlin, H., Carucci, D.J., Cummings, L.M., Aravind, L., Koonin, E.V., Shallom, S., Mason, T., Yu, K., Fujii, C., et al. 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282:** 1126–1132.
Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419:** 498–511.
Hall, T.M., Porter, J.A., Young, K.E., Koonin, E.V., Beachy, P.A., and Leahy, D.J. 1997. Crystal structure of a Hedgehog autoprocessing domain: Homology between Hedgehog and self-splicing proteins. *Cell* **91:** 85–97.
Hasegawa, M., Kishino, H., and Saitou, N. 1991. On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol*. **32:** 443–445.
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.
Kumar, S., Tamura, K., Jakobsen, I.B., and Nei, M. 2001. MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244–1245.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Laybourn-Parry, J. 1984. *A functional biology of the free-living protozoa*. University of California Press, Berkeley, CA.
Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J., et al. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301:** 1503–1508.
Lespinet, O., Wolf, Y.I., Koonin, E.V., and Aravind, L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res*. **12:** 1048–1059.
Mann, T. and Beckers, C. 2001. Characterization of the subpellicular network, a filamentous membrane skeletal component in the parasite *Toxoplasma gondii*. *Mol. Biochem. Parasitol*. **115:** 257–268.
Matussek, K., Moritz, P., Brunner, N., Eckerskorn, C., and Hensel, R. 1998. Cloning, sequencing, and expression of the gene encoding cyclic 2,3-diphosphoglycerate synthetase, the key enzyme of cyclic 2,3-diphosphoglycerate metabolism in *Methanothermus fervidus*. *J. Bacteriol*. **180:** 5997–6004.
Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*. **10:** 1–6.
Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*. **302:** 205–217.
Odenthal-Schnittler, M., Tomavo, S., Becker, D., Dubremetz, J.F., and Schwarz, R.T. 1993. Evidence of *N*-linked glycosylation in *Toxoplasma gondii*. *Biochem. J*. **291:** 713–721.
Peterson, D.S., Miller, L.H., and Wellems, T.E. 1995. Isolation of multiple sequences from the *Plasmodium falciparum* genome that encode conserved domains homologous to those in erythrocyte-binding proteins. *Proc. Natl. Acad. Sci*. **92:** 7100–7104.
Pollack, Y., Kogan, N., and Golenser, J. 1991. *Plasmodium falciparum*: Evidence for a DNA methylation pattern. *Exp. Parasitol*. **72:** 339–344.
Ponting, C.P., Aravind, L., Schultz, J., Bork, P., and Koonin, E.V. 1999. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol*. **289:** 729–745.
Pradel, G., Hayton, K., Aravind, L., Iyer, L.M., Abrahamsen, M.S., Bonawitz, A., Mejia, C., and Templeton, T.J. 2004. A multidomain adhesion protein family expressed in *Plasmodium falciparum* is essential for transmission to the mosquito. *J. Exp. Med*. **199:** 1533–1544.
Reeves, P.R., Hobbs, M., Valvano, M.A., Skurnik, M., Whitfield, C., Coplin, D., Kido, N., Klena, J., Maskell, D., Raetz, C.R., et al. 1996. Bacterial polysaccharide synthesis and gene nomenclature. *Trends Microbiol*. **4:** 495–503.
Ritte, G., Lloyd, J.R., Eckermann, N., Rottmann, A., Kossmann, J., and Steup, M. 2002. The starch-related R1 protein is an α-glucan, water dikinase. *Proc. Natl. Acad. Sci*. **99:** 7166–7171.
Ronquist, F. and Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.
Rost, B., Fariselli, P., and Casadio, R. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci*. **5:** 1704–1718.
Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. 2002. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18:** 502–504.
Smith, J.D., Chitnis, C.E., Craig, A.G., Roberts, D.J., Hudson-Taylor, D.E., Peterson, D.S., Pinches, R., Newbold, C.I., and Miller, L.H. 1995. Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82:** 101–110.
Su, X.Z., Heatwole, V.M., Wertheimer, S.P., Guinet, F., Herrfeldt, J.A., Peterson, D.S., Ravetch, J.A., and Wellems, T.E. 1995. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* **82:** 89–100.
Subramanian, G., Koonin, E.V., and Aravind, L. 2000. Comparative genome analysis of the pathogenic spirochetes *Borrelia burgdorferi* and *Treponema pallidum*. *Infect. Immun*. **68:** 1633–1648.
Tang, L.Y., Reddy, M.N., Rasheva, V., Lee, T.L., Lin, M.J., Hung, M.S., and Shen, C.K. 2003. The eukaryotic DNMT2 genes encode a new class of cytosine-5 DNA methyltransferases. *J. Biol. Chem*. **278:** 33613–33616.
Templeton, T.J., Lancto, C.A., Vigdorovich, V., Liu, C., London, N.R., Hadsell, K.Z., and Abrahamsen, M.S. 2004. The *Cryptosporidium* oocyst wall protein is a member of a multigene family and has a homolog in *Toxoplasma*. *Infect. Immun*. **72:** 980–987.
Varki, A., Cummings, R., Esko, J., Freeze, H., Hart, G., and Marth, J. 1999. *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Zhu, G., Keithly, J.S., and Philippe, H. 2000. What is the phylogenetic position of *Cryptosporidium*? *Int. J. Syst. Evol. Microbiol*. **50:** 1673–1681.

WEB SITE REFERENCES

ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.txt; BLASTCLUST.
http://134.84.110.219/cgi-bin/gbrowse/crypto909; *C. parvum* genome sequence information and annotation.
http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html; TOPRED1.0.
http://www.cbs.dtu.dk/services/SignalP-2.0/; SIGNALP.
http://www.cryptodb.org; *C. parvum* genome sequence information and annotation.
http://www.ncbi.nlm.nih.gov/CBBresearch/Walker/SEALS/index.html; SEALS.
http://www.sanger.ac.uk/Software/Pfam/index.shtml; PFAM database.
http://www.tree-puzzle.de; TreePuzzle 4.02.
http://biowulf.nih.gov; Biowulf processing system at the National Institutes of Health.

Received March 23, 2004; accepted in revised form June 14, 2004.