IDENTIFICATION AND FUNCTIONAL CHARACTERIZATION OF *CIS*-REGULATORY ELEMENTS IN APICOMPLEXAN PARASITES

by

NANDITA MULLAPUDI

(Under the Direction of JESSICA C KISSINGER)

ABSTRACT

Apicomplexan parasites, a group of phylogenetically divergent protist parasites are known for causing several diseases including malaria, cryptosporidiosis and toxoplasmosis. Gene-regulatory mechanisms in this group of parasites remain largely unknown. Owing to their medical and evolutionary importance, several apicomplexan genomes have been sequenced. This thesis describes an effort to identify and characterize novel cis-regulatory elements in the genomes of *Toxoplasma gondii* and *Cryptosporidium parvum* by using a combination of computational and experimental tools. I have shown that cis-regulatory elements play a significant role in the regulation of gene expression in these parasites, as is evidenced by the presence of conserved cis-motifs that are associated with gene expression.

INDEX WORDS: Apicomplexa, MEME, *Toxoplasma gondii*, *Cryptosporidium parvum*, transcription, *cis*-regulatory

IDENTIFICATION AND FUNCTIONAL CHARACTERIZATION OF CIS-REGULATORY ELEMENTS IN APICOMPLEXAN PARASITES

by

NANDITA MULLAPUDI

M.Sc., UNIVERSITY OF MUMBAI, INDIA, 2001

B.Sc., UNIVERSITY OF MUMBAI, INDIA, 1999

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

MAY 2007

IDENTIFICATION AND FUNCTIONAL CHARACTERIZATION OF CIS-REGULATORY ELEMENTS IN APICOMPLEXAN PARASITES

by

NANDITA MULLAPUDI

Major Professor: JESSICA C. KISSINGER

Committee: JEFFREY BENNETZEN
ROBERT IVARIE
DAVID PETERSON
BORIS STRIEPEN

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2007

DEDICATION

Dedicated to my mother Sakuntala and my father M.S.V.Rao for being my unending sources of wisdom and inspiration.

ACKNOWLEDGEMENTS

First and foremost I'd like to thank Jessie for being an excellent mentor and imparting many valuable lessons over the past five years. I want to thank her for allowing me independence in the choice and workings of my project, while constantly reminding me to think critically and stay focused on the big picture. I am thankful for her patient efforts in honing my skills at scientific writing and presentation, and critical comments on my writing that greatly improved the quality of this dissertation. Additionally, she has been very supportive during times of personal crises and her help at such times is much appreciated. My committee members: Dr. Boris Striepen, Dr. David Peterson, Dr. Bob Ivarie and Dr. John McDonald have been of great help right from the design of my project until it's completion. They were always available for advice beyond the mandatory committee meetings and I am grateful to them for their support and time. I'd like to thank Dr. Jeff Bennetzen for agreeing to be on my thesis committee with very short notice.

Past and present members of the Kissinger lab have been of great help in various ways. I'd like to thank the database group for making my computational research easy and catering to my requests from time to time. Madeline Cozad, Libby Mathis, Karen Hermetz and Abhijeet provided invaluable help in the wet lab. Adriana, Abhijeet, Erica and Kathy have not only been ever-helpful colleagues but also wonderful friends who made lab-life more fun. Past and current members of the Striepen lab have also been extremely helpful and great people to work with. I particularly wish to thank Marc Jan Gubbels for teaching me the ropes with parasite and cell-culture.

The department of Genetics and the CTEGD have provided a scientifically inspiring, warm and friendly atmosphere to work in. The McEachern lab, the Tarleton lab and the DoCampo & Moreno lab have been great lab-neighbors, always willing to share resources and expertise. Tammy Andros, Kathy Couch, Erica Young, Janice Lunsford and Larry Peters have always been extremely helpful in administrative matters and helped facilitate science as smoothly as possible. Dave Brown has been a great buddy and always came to my rescue in times of computer and poster crises.

I'd like to thank my friend Ramneek Gupta for introducing me to the field of bioinformatics. My friends Priyanka, Ratna, Shira, Hili, Itamar, Madhavi, Minnie, Chenoo, Bhargavi, Tina, Malika and Sonu were my regular sources of good times, inspiration and support through the ups and downs of research and life. I am grateful to Saurabh for his unfailing daily wake-up call service that saw me through my qualifiers, to Shira and Kaushik for motivating me and ensuring that this thesis got written and to Karen and Kathy for help with last minute crisis control.

Several teachers from school and my undergraduate years at S.I.E.S College nurtured my aptitude for science and encouraged me to pursue my goals. My parents taught me the importance of logical thinking, perseverance and hard work. Their love and constant encouragement are why I am here today. My sister Savita has always been there for me and along with my brother-in-law Hari kept my spirits up and have been my home away from home. Finally, when all else failed, my little niece Aria always succeeded in making me smile. ☺

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

The phylum apicomplexa consists of over 4000 species, all of which are obligate intracellular protozoan parasites (Levine 1988). Members of this phylum cause diseases of human and veterinary importance and examples include *Plasmodium* (the causative agent of malaria), *Cryptosporidium* (the causative agent of cryptosporidiosis) and *Toxoplasma* (the causative agent of toxoplasmosis). Malaria is responsible for around 2.7 million deaths worldwide every year (WHO report, 2000) and *Toxoplasma* and *Cryptosporidium* are gaining increasing notoriety as AIDS-related pathogens, responsible for chronic and morbid infections in immuno-compromised individuals (Kasper and Buzoni-Gatel 1998). The Apicomplexa belong to the super-group alveolata, and like the other protists, display a complex evolutionary history and as yet unresolved phylogenetic relationships with other eukaryotes (Escalante and Ayala 1995). While studies of fundamental biological processes in the animals, higher plants as well as in fungi have benefited from the existence of (to some extent) generalized methods and availability of model systems to elucidate some of the basic underpinnings of molecular genetics, there is no general protistan model organism. The protists encompass a wide range of unicellular eukaryotes, reflecting a rich diversity and unknown relationships on the tree of life (Baldauf 2003). As more genomes are sequenced, the protists are making an appearance on every branch of the tree of life, exhibiting more divergence from one another than that exhibited between plants and animals (Baldauf 2003; Adl et al. 2005). This wide diversity is apparent in the different life-styles as well as the genetic and cellular organization in the different protists

(Parfrey et al. 2006).

Within the phylum Apicomplexa, all members are parasitic. They are characterized by compact genomes (9 Mb to 65 Mb), obligate intracellular lifestyles and complex developmental cycles (Ajioka 1997), (Figure 2.2- 2.4). Genome evolution in the Apicomplexa has been shaped by a myriad of forces such as horizontal gene transfer events (Huang et al. 2004b; Huang et al. 2004a), and the retention of a relict plastid-like organelle, the apicoplast, which originated as a result of a secondary endosymbiotic event in an ancestral apicomplexan (Delwiche 1999). The apicoplast is a non-photosythetic organelle essential for survival in many apicomplexans (*T. gondii* and *P. falciparum*) but has been lost in the case of the basal apicomplexan *C. parvum* (Zhu et al. 2000). The apicomplexa alone are believed to encompass several hundred million years of divergence (Escalante and Ayala 1995); much greater than the divergence within the Vertebrata or within the different fungal species. The exact estimate in mya is debatable and likely an overestimate, but nevertheless the diversity of the members of this phylum is quite apparent in the considerable variation seen within different members of this phylum in terms of life-style, host-preference and genome organization (Gardner et al. 2002; Abrahamsen et al. 2004). The differences relevant to the work described in this thesis are described here for three apicomplexan species: *P. falciparum*, *T. gondii* and *C. parvum*.

1.1 Life-cycle and gene expression:

1.11 *Plasmodium falciparum:*

P. falciparum, a haemosporidian parasite carries out its life-cycle utilizing two hosts, the *Anopheles* mosquito (sexual phase) and the human host (asexual phase) (Figure 1.2). During a mosquito bite, sporozoites are released from the mosquito which first invade and develop into schizonts in the liver cells of the human host (the extra-erythrocytic stage). The schizonts then

transform and release merozoites into the bloodstream (the intra-erythrocytic cycle). Merozoites invade red blood cells to undergoing an additional round of multiplication to produce 12-16 merozoites contained within a schizont. Some merozoites can differentiate into sexual forms-gametocytes, that are taken up a mosquito during a blood meal. Within the mosquito mid-gut, ookinetes are formed from the fertilization of gametocytes. The ookinete traverses the mosquito gut wall and encysts at the exterior as an oocyst. Soon the oocyst ruptures, releasing hundreds of sporozoites into the mosquito body cavity from where it can be transmitted to a vertebrate host. Malaria is thus transmitted via mosquito bites; the sexual part of the life-cycle is essential for the propagation of the parasite. Analyses of mRNA and protein levels of the parasite from these different life-cycle stages has revealed that the parasite exhibits a tightly controlled cascade of gene expression, possibly triggered by a single developmental induction signal as is observed in the sequential expression of stage specific transcripts throughout it's life-cycle (Le Roch et al. 2004).

1.12 *Toxoplasma gondii*:

The coccidian parasite *T. gondii* also carries out its life-cycle through two hosts: the feline host acts as the definitive host for the sexual phase of its life-cycle, and practically any warm-blooded animal can serve as the intermediate host during its asexual phase (Ajioka 1997) (Figure 1.3). *Toxoplasma gondii* is unique in its ability to propagate solely through the asexual phase, and to infect such a wide host-range therein. This unique property is now thought to be the result of a single meiotic event that gave rise to a population of parasites (~10,000 years ago) endowed with the useful adaptation of direct oral infectivity, allowing it to bypass the sexual phase in the cat and propagate solely asexually (Su et al. 2003; Boyle et al. 2006). Within its asexual phase, the parasite can adopt a rapidly dividing form, the tachyzoite that gives rise to a

quiescent form known as the bradyzoite. The developmental triggers for the parasite to switch between these forms is an area of active investigation, and most gene expression studies in this parasite have been conducted on the expression of stage-specific genes (Manger et al. 1998). More recently, a large-scale serial analysis of gene expression (SAGE) study of the different life-cycle stages of the parasite has shown that the *T. gondii* transcriptome is highly dynamic, with stage-specific populations of mRNAs being expressed at different stages and 20% of the mRNA pool consisting of apicomplexan-specific messages (Radke et al. 2005).

1.13 *Cryptosporidium parvum*:

Cryptosporidium parvum is a basal apicomplexan parasite of uncertain affinity to the other apicomplexans (Xiao et al. 2004). Its entire life-cycle takes place within the intestinal tract of the same host (Figure 2.4) in contrast to the heteroxenous life-cycle (utilizing more than one host) seen in the case of *P. falciparum* and *T. gondii* above. Oocysts of *C. parvum* released in the feces of infected hosts are capable of infecting new hosts via the fecal-oral route. These oocysts release sporozoites that attach to the intestinal epithelial cells where they develop into trophozoites, undergo asexual reproduction and produce merozoites that are released into the intestinal lumen. These merozoites can infect new intestinal epithelial cells, or, alternatively, develop into macro or micro-gametocytes after infecting an enterocyte. The resulting zygote (formed from the fertilization of the gametocytes) can develop into thick walled oocysts that will exit the host, or thin-walled oocysts capable of auto-infection. Due to the lack of an efficient *in vitro* culture system for propagating *C. parvum*, the oocysts and sporozoites are the best sources of parasitic material for molecular and biochemical analyses (Spano and Crisanti 2000). More recently, comparative RT-PCR techniques have been developed to study parasite gene expression within a 72h time-window after infecting HCT-8 cells in tissue culture with oocysts

(Abrahamsen and Schroeder 1999). These studies enable us to assess gene expression trends within a small window of the life-cycle, and has led to the discovery of developmentally regulated genes (Templeton et al. 2004; Mullapudi et al. 2007).

1.2 Nuclear genome organization

Apicomplexan genomes vary in size and composition: *C. parvum* has the smallest genome at ~9.1 Mb and 8 chromosomes, *P. falciparum* has a 23 Mb genome and 14 chromosomes, and *T. gondii* has the largest genome of ~ 64 Mb, and 12 chromosomes. The nuclear genomes for all three of these parasites show mono*cis*tronic gene-organization, and the genes of both *P. falciparum* and *T. gondii* contain introns. In *C. parvum* less than 20% of the genes are thought to contain introns. Co-expressed genes do not cluster in terms of their genomic location, with the exception of some gene families (example: the *LSP* gene family in *C. parvum*). *Plasmodium falciparum* and *C. parvum* genomes show a strong AT bias (80% AT and 70%AT respectively), while *T. gondii* has a more uniform composition (45% AT) (Gardner et al. 2002; Abrahamsen et al. 2004).

1.3 Summary

The phylum Apicomplexa comprises of a diverse group of parasites that present an intriguing mystery in terms of their gene-regulatory machinery. Preliminary gene-specific studies have revealed the absence of canonical gene-regulatory elements in this group of parasites. Whole-genome sequence analyses and large-scale experiments (which were unavailable at the initiation of this study) have provided some clues about gene expression patterns. Given their divergence and evolutionary complexity, it is likely that this group of parasites possesses a very divergent regulatory mechanism not yet seen in other systems. On the other hand, some novel molecular phenomena such as histone-mediated regulation, the discovery

of self-splicing RNA and the discovery of telomerase were first discovered and thought to be unique in the ciliate *Tetrahymena thermophila* (Collins and Gorovsky 2005), and then found to be used widely in other organisms. It will be interesting to determine how apicomplexa regulate their genes and how much of their genetic repertoire is dedicated to different levels of genetic control. Given the significant differences within the Apicomplexa, it is also conceivable that they will not all follow the same modes of gene regulation to the same extent.

The obligate intracellular nature of the parasites and their close relationship with the host cell makes it complicated to isolate and study parasite-specific functions in this group of parasites, providing one with limited tools to investigate fundamental aspects such as gene regulation. On the other hand, several apicomplexan parasites have had their genomes sequenced, owing to their medical and evolutionary importance. My goal has been to better understand how these parasites regulate their genes, and we have made use of the vast resource of sequence data and available computational tools to uncover clues about transcriptional gene regulation in two of these parasites, *T. gondii* and *C. parvum*. Our methods present a vast improvement over traditional promoter-bashing experiments for the initial detection of putative regulatory elements and their targeted characterization when possible. I describe these methods and findings in the subsequent chapters of this thesis.

1.4 Organization of this thesis

This thesis is organized into 5 chapters. In Chapter 2, I present the background for this thesis in two parts. In section 2.1, I have conducted a survey of the literature on what is known about gene regulation in the apicomplexans *P. falciparum* and *T. gondii*. In section 2.2, I describe an introduction to computational approaches to *cis*-element detection as is relevant to this study. In Chapter 3, I present a study on *cis*-element detection in *Cryptosporidium parvum*.

We report the presence of putative *cis*-regulatory elements identified upstream of genes that show concerted expression. In Chapter 4, I describe a study on *cis*-element detection in *Toxoplasma gondii*. Here we report the presence of *cis*-regulatory elements and also functionally characterize their role in gene regulation through experimental methods. Finally, in Chapter 5, I discuss and present ideas for future direction of the work described in this thesis.

References

Abrahamsen MS, Schroeder AA (1999) Characterization of intracellular *Cryptosporidium parvum* gene expression. Mol Biochem Parasitol 104(1): 141-146.

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G et al. (2004) Complete Genome Sequence of the Apicomplexan, *Cryptosporidium parvum*. Science 304: 441-445.

Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR et al. (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 52(5): 399-451.

Ajioka JW (1997) The protozoan phylum Apicomplexa. Methods 13(2): 79-80.

Baldauf SL (2003) The deep roots of eukaryotes. Science 300(5626): 1703-1706.

Boyle JP, Rajasekar B, Saeij JP, Ajioka JW, Berriman M et al. (2006) Just one cross appears capable of dramatically altering the population biology of a eukaryotic pathogen like *Toxoplasma gondii*. Proc Natl Acad Sci U S A 103(27): 10514-10519.

Collins K, Gorovsky MA (2005) *Tetrahymena thermophila.* Curr Biol 15(9): R317-318.

Delwiche CF (1999) Tracing the Thread of Plastid Diversity through the Tapestry of Life. Am Nat 154(S4): S164-S177.

Escalante AA, Ayala FJ (1995) Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes. Proc Natl Acad Sci U S A 92(13): 5793-5797.

Gardner MJ, Hall N, Fung E, White O, Berriman M et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419(6906): 498-511.

Huang J, Mullapudi N, Sicheritz-Ponten T, Kissinger JC (2004a) A first glimpse into the pattern and scale of gene transfer in Apicomplexa. International journal for parasitology 34(3): 265-274.

Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS et al. (2004b) Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. Genome Biol 5(11): R88.

Kasper LH, Buzoni-Gatel D (1998) Some Opportunistic Parasitic Infections in AIDS: Candidiasis, Pneumocystosis, Cryptosporidiosis, Toxoplasmosis. Parasitology today (Personal ed 14(4): 150-156.

Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A et al. (2004) Global analysis of transcript and protein levels across the *Plasmodium falciparum* life-cycle. Genome Res 14(11): 2308-2318.

Levine ND (1988) Progress in taxonomy of the Apicomplexan protozoa. J Protozool 35(4): 518-520.

Manger ID, Hehl A, Parmley S, Sibley LD, Marra M et al. (1998) Expressed sequence tag analysis of the bradyzoite stage of *Toxoplasma gondii*: identification of developmentally regulated genes. Infect Immun 66(4): 1632-1637.

Mullapudi N, Lancto CA, Abrahamsen MS, Kissinger JC (2007) Identification of putative *cis*-regulatory elements in *Cryptosporidium parvum* by *de novo* pattern finding. BMC Genomics 8: 13.

Parfrey LW, Barbero E, Lasser E, Dunthorn M, Bhattacharya D et al. (2006) Evaluating Support for the Current Classification of Eukaryotic Diversity. PLoS Genet 2(12): e220.

Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS et al. (2005) The transcriptome of *Toxoplasma gondii*. BMC Biol 3: 26.

Spano F, Crisanti A (2000) *Cryptosporidium parvum*: the many secrets of a small genome. International journal for parasitology 30(4): 553-565.

Su C, Evans D, Cole RH, Kissinger JC, Ajioka JW et al. (2003) Recent expansion of *Toxoplasma* through enhanced oral transmission. Science 299(5605): 414-416.

Templeton TJ, Lancto CA, Vigdorovich V, Liu C, London NR et al. (2004) The *Cryptosporidium* oocyst wall protein is a member of a multigene family and has a homolog in *Toxoplasma*. Infect Immun 72(2): 980-987.

Xiao L, Fayer R, Ryan U, Upton SJ (2004) *Cryptosporidium* taxonomy: recent advances and implications for public health. Clin Microbiol Rev 17(1): 72-97.

Zhu G, Marchewka MJ, Keithly JS (2000) *Cryptosporidium parvum* appears to lack a plastid genome. Microbiol Mol Biol Rev 146: 315-321.

Figure 1.1 Current view of the tree of life: Positions of the super groups within the eukaryotes revealed by multiple protein-encoding gene phylogenies

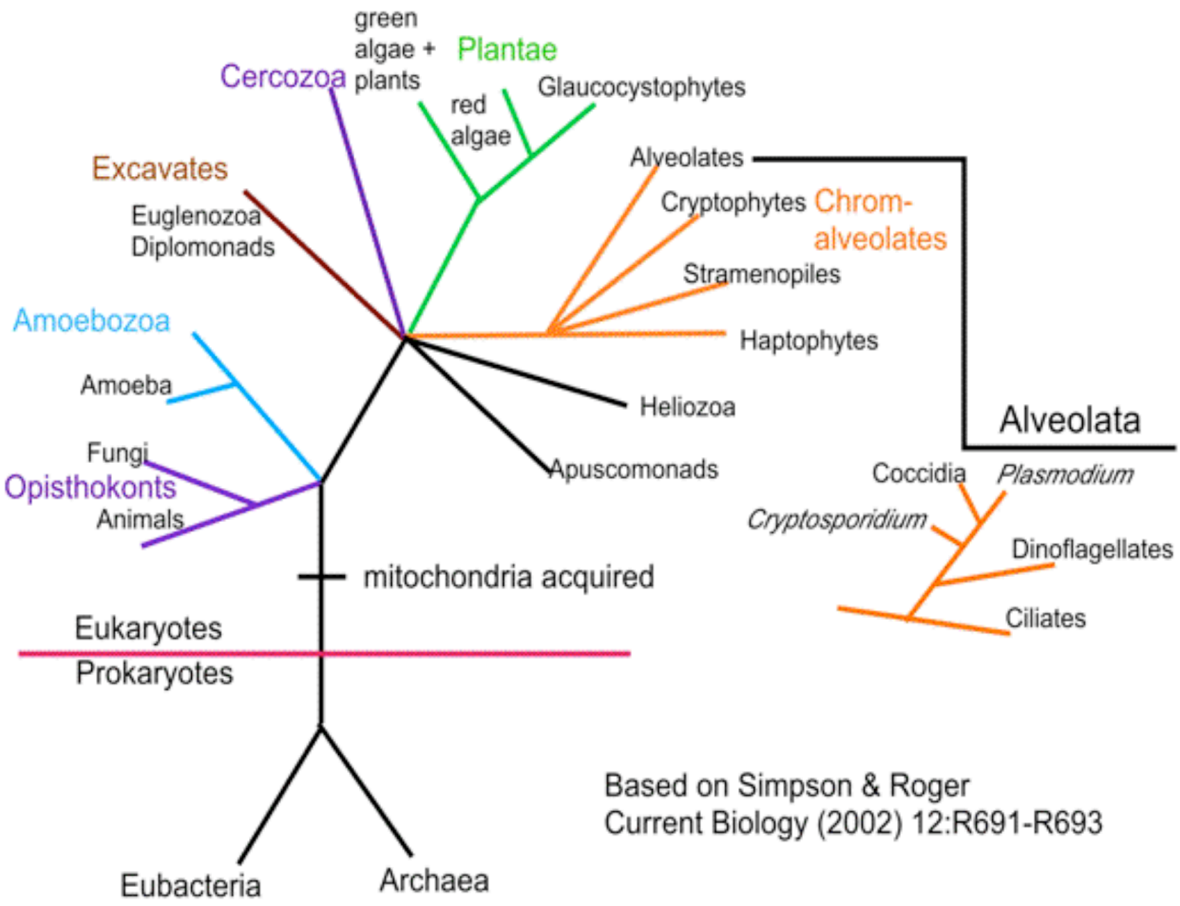


Figure 1.1

Current view of the tree of life

Figure 1.2 Life-cycle of *Plasmodium falciparum*: An infected mosquito can transfer sporozoites into the blood stream of the human host, which then travel to the liver. The parasite replicates in the liver, and merozoites are released in the bloodstream. Merozoites bind to the surface of the red blood cell (RBC) and enter the RBC where they undergoes growth through the ring and trophozoite stages. This results in the production of schizonts containing multiple merozoites (erythrogenic / intra-erythrocytic cycle). Mature schizonts destroy the RBCs and release merozoites into the bloodstream, which invade new RBCs. Occasionally, parasite maturation can cause the production of gametocytes which are released into the bloodstream and are subsequently taken up by the mosquito, via a blood meal, and then enter the sexual stage of development (sporogyny).

(Lamb et al, Expert Reviews in Molecular Medicine, 8:6 2006)

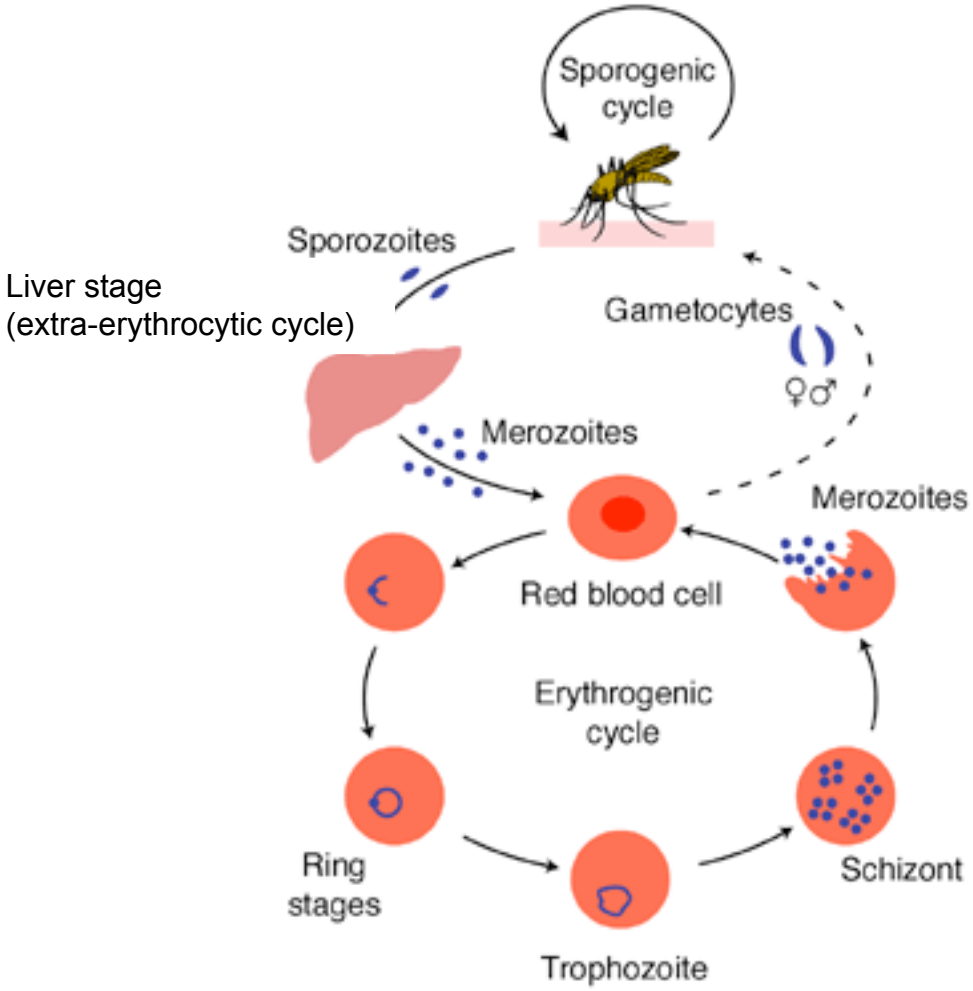


Figure 1.2

Life-cycle of *Plasmodium falciparum*

(Lamb et al, Expert Reviews in Molecular Medicine, 8:6 2006)

Figure 1.3 Life-cycle of *Toxoplasma gondii* : The life-cycle includes both sexual and asexual modes of proliferation and transmission. The sexual cycle can only take place in members of the cat family (Felidae). (a, b) After the cat ingests tissue cysts, the parasites invade the enterocytes,, divide and (c) differentiate into microgametocytes and macrogametocytes. (d) The gametocytes fuse to form a zygote, which upon meiotic division produce oocysts that are shed into the environment with the cat's faeces. (e) The oocyst divides by meiosis and produces highly infectious 'sporozoites' that can persist for years in a moist environment. (f) After ingestion (by a secondary host such as a mouse), (g) sporozoites differentiate into the rapidly dividing 'tachyzoite' form, which establishes and sustains the acute infection. (h) During the acute infection, congenital transmission to the developing fetus can occur. (i) Sometimes, the tachyzoite changes into a slowly dividing form known as the 'bradyzoite'. Latent bradyzoite tissue cysts persist for the life of the host, re-emerging occasionally, but do not produce overt disease in healthy individuals. (j) Carnivorous ingestion of tissue cysts can lead to the infection of a naive host, allowing for an indefinite nonsexual propagation of *T. gondii*. (k) In the cat, this will initiate the sexual cycle. The solid lines indicate parasite differentiation and the dashed lines indicate modes of transmission (Ajioka et al, 2001).

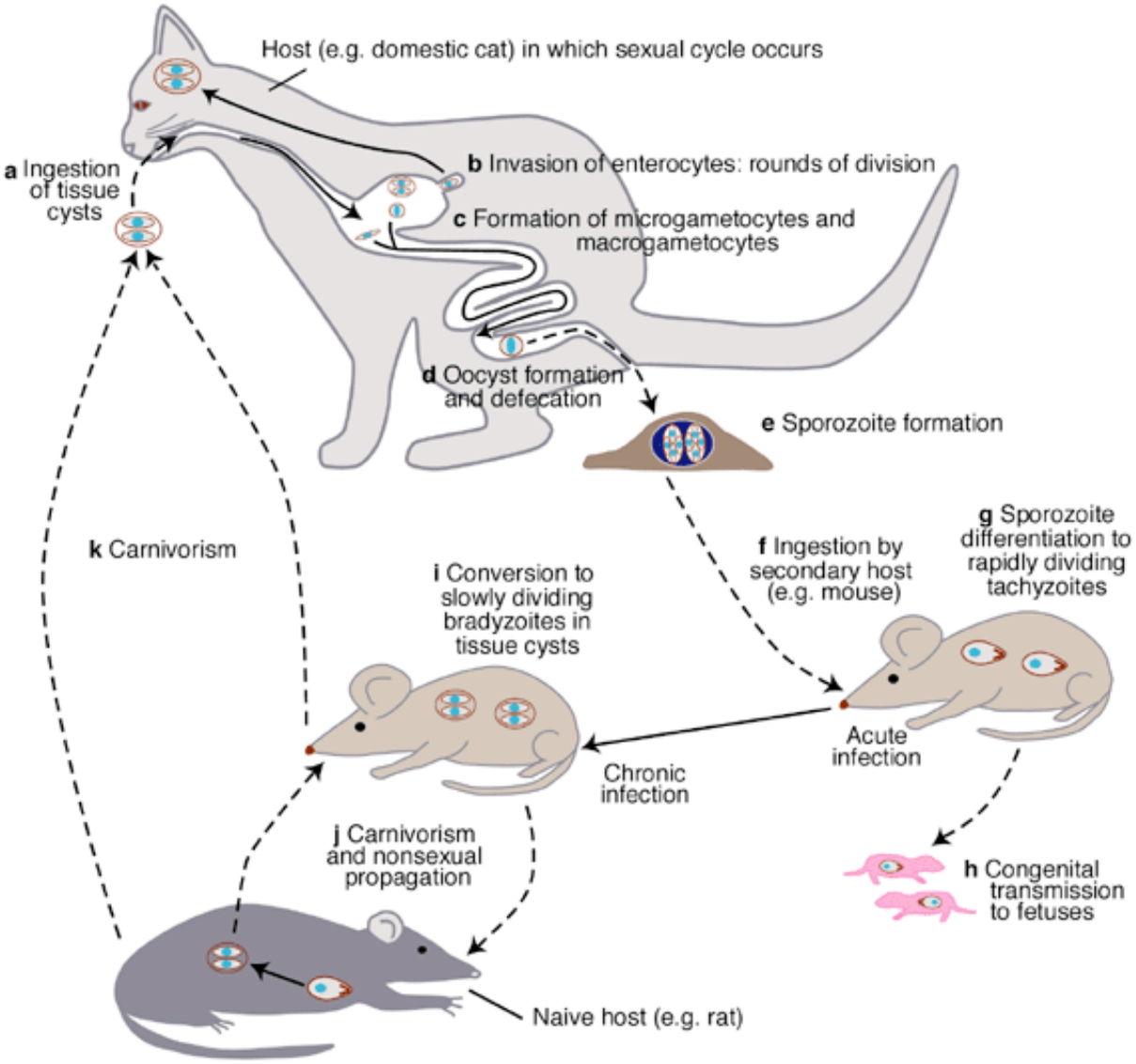


Figure 1.3

Life-cycle of *Toxoplasma gondii*

(Ajioka et al, Expert Reviews in Molecular Medicine 2001)

Figure 1.4 Life-cycle of *Cryptosporidium parvum:* Oocysts of *C. parvum* released in the feces of infected hosts are capable of infecting new hosts via the fecal-oral route. These oocysts release sporozoites that invade and attach to the intestinal epithelial cells where they develop into trophozoites, undergo asexual reproduction and produce merozoites that are released into the intestinal lumen. These merozoites can infect new intestinal epithelial cells, or, alternatively, develop into macro or micro-gametocytes after infecting an enterocyte. The resulting zygote (formed from the fusion of the gametocytes) then undergoes meiosis to release thin-walled or thick-walled oocysts. Thick-walled oocysts are excreted in the feces. Thin walled oocysts can also excyst within the host intestinal tract and cause autoinfection.

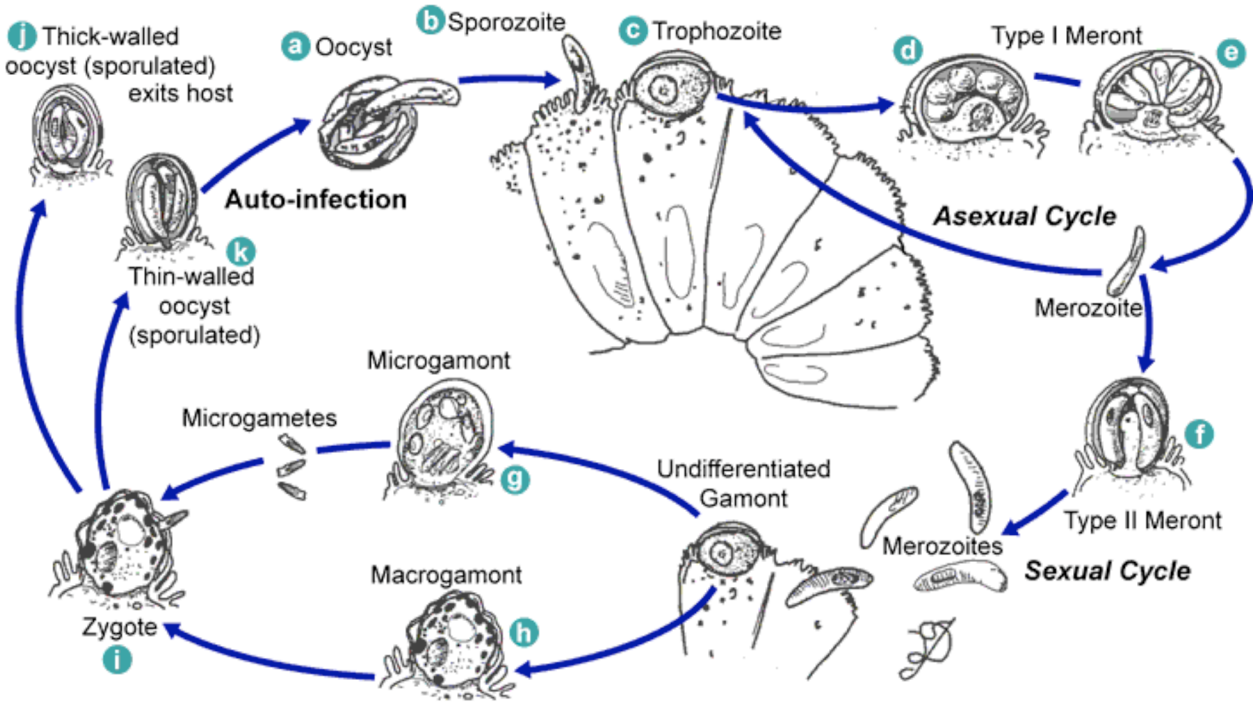


Figure 1.4

Life-cycle of *Cryptosporidium parvum*

(Adapted from http://www.dpd.cdc.gov/)

CHAPTER 2

BACKGROUND

This section is comprised of two parts. Section 2.1 is a survey of the literature on apicomplexan gene regulation for the parasites *T. gondii*, *P. falciparum* and *C. parvum*. Section 2.2 provides an introduction to computational methods of detecting *cis*-regulatory elements as is relevant to the work described in this thesis.

2.1 Gene Regulation in the Apicomplexa (Review Of The Literature)

All members of the phylum Apicomplexa are characterized by complex developmental lifecycles and obligatory parasitic lifestyles, evidencing a need for the regulation of gene expression at various levels to facilitate processes such as invasion, immune evasion, establishment of infection and developmental stage-transition (Perkins et al. 2000). In eukaryotic cells, gene expression is controlled at various levels. Briefly, gene regulation can take place at the chromatin level via histone modifications that render chromatin in an active or inactive state; at the transcriptional level, mediated by basal and specific transcription factors and/or repressors and at the post-transcriptional level by the regulation of transcript levels, message stability and RNA-mediated gene silencing (van Driel et al. 2003). In section 2.1.1, I discuss whole-genome surveys and large-scale analyses that provide clues about the general trends of gene regulation in the Apicomplexa and in section 2.1.2, I discuss gene-specific studies and experimental evidence for the various levels of regulation as described above.

2.1.1 Insights into apicomplexan gene regulation: clues from genome-wide studies in *P. falciparum*, *T. gondii* and *C. parvum*

Very little is known about gene regulation in the apicomplexans. Early studies involved gene-by-gene examination of promoters and provided early clues about gene-regulatory mechanisms in these parasites (Soldati and Boothroyd 1995), (Bohne et al. 1997). The availability of whole-genome sequences gave rise to alternative means to study gene-regulatory signals and trends and examine the repertoire of regulatory machinery. The genome sequence of the malarial parasite *P. falciparum* revealed a paucity of known transcription factors encoded in the genome, and a large percentage of "hypothetical proteins" for which no function could be assigned based on sequence similarity (Gardner et al. 2002). Several large-scale microarray experiments were performed to characterize the *P. falciparum* transcriptome. In one study, a spotted oligonucleotide array was employed to study transcriptional changes within the intraerythrocytic developmental cycle of the malarial parasite (Bozdech et al. 2003). A separate study (Le Roch et al. 2003) included mRNA from the intra-erythrocytic stage and two additional stages, the sporozoites and the gametocytes and analyzed their expression profiles. Both studies revealed a similar and a striking trend of gene expression in *P. falciparum*. It was found that groups of genes peaked in expression only once per life-cycle, exactly at the time that they were required, as was inferred from their function and the corresponding developmental stage. For example, trophozoites expressed high levels of genes involved in DNA replication; schizonts- the stage that precedes invasive merozoites expressed high levels of genes involved in invasion. The coordinate expression of genes in this manner in most cases could not be ascribed to contiguity of genes at the chromosomal level. Subsequent analyses of the *P. falciparum* proteome in relation to the transcriptome (Le Roch et al. 2004) showed that while a strong

correlation existed between the production of an mRNA and its corresponding protein, there was typically a delay between the maximal detection of a transcript and the appearance of its corresponding protein, so that the transcriptome of a particular stage correlated best with the proteome of the following stage. These data indicated that post-transcriptional regulatory mechanisms could be responsible for control of gene expression in *P. falciparum*.

To investigate the role of conserved upstream sequences motifs in stage specific gene expression (van Noort and Huynen 2006), Van Noort and colleagues clustered genes in *P. falciparum* that showed significant correlation in their expression profiles and retrieved their corresponding orthologs in *P. yoelii*. They detected 79 over-represented sequence motifs conserved in both species in all the upstream regions (5334 genes), of which 12 motifs correlated with specific expression patterns. They were able to establish a strong correlation between the presence of a conserved motif and stage-specific gene expression for genes upregulated in the schizont stage, ring stage and in the gametocytes, indicating the role of specific sequence motifs in stage-specific gene expression. It would be interesting to see if the same genes also showed concordant protein levels, which would tell us the extent to which the sequence-specific signals were important for their expression. They hypothesize that the small number of total regulatory motifs indicates that *P. falciparum* employs a combinatorial mode of gene regulation, making use of different combinations of the small number (12) of motifs to establish different expression patterns. It is not proven beyond doubt that these are the only regulatory motifs in the genome; other motifs could be subtle and not statistically over-represented but still have biological function; still others could be species-specific and not be identified in a cross-species study. The biological function of the twelve motifs was not established, but the strong correlation with co-

expressed genes is indeed encouraging and this study serves an excellent starting point to further elucidate mechanisms employed by *P. falciparum* to regulate gene expression.

Surveys of the *T. gondii* genome have also revealed a paucity of known transcription factors. While most of the machinery required for general pol II transcriptional control have been identified (Meissner and Soldati 2005), specialized transcription factors as seen in other eukaryotes are not found here. In *T.gondii*, studies have focused on the analysis of gene expression during stage-switching in the asexual phase of the parasite life-cycle. Analysis of ESTs from bradyzoites and tachyzoites, the two asexual stages, revealed the presence of several sequence clusters specific to either developmental stage (Manger et al. 1998). *T. gondii* tachyzoites are capable of switching to the bradyzoite stage in presence of specific stresses in vitro such as γ-interferon treatment, high pH and high temperature thus mimicking host-response in an infected cell (Cleary et al. 2002). A time-course based expression profile of parasites undergoing *in vitro* differentiation to bradyzoites was compared to tachyzoites by microarray analysis (Cleary et al. 2002) and 31 genes were reported to be preferentially induced in the bradyzoite stage. Mutant parasites incapable of tachyzoite to bradyzoite differentiation (Tbd mutants) showed lowered expression of a subset (48%) of the 31 aforementioned bradyzoite specific genes. Although limited in scope due to the small number of genes, these studies have pointed to the presence of transcriptional control of stage-specific gene expression, and set the framework for a hierarchical model of gene regulation associated with *T. gondii* development. In the first large-scale investigation of the *T. gondii* transcriptome, serial analysis of gene expression (SAGE) was employed to measure transcript abundance across key developmental stages of the parasite life-cycle (Radke et al. 2005). Stage-specific libraries were constructed from oocysts containing mature sporozoites, parasites emerging on day 4 post-sporozoite

infection, rapidly growing tachyzoites on day 6 post-inoculation, parasites immediately following growth shift at day 7 post-inoculation and from a mixed population of slowly growing tachyzoites and bradyzoites at day 15 post-inoculation. From these analyses, it was found that groups of co-regulated mRNAs were associated with the major developmental transitions. Large numbers of mRNAs were expressed exclusively in a single developmental stage, similar to the trend observed in *P. falciparum*. Genes that showed co-regulation were scattered throughout the genome, pointing to gene-specific methods of regulation for their expression. Finally, it was seen that 20% of the developmentally regulated transcripts were apicomplexan-specific with no known orthologs outside the Apicomplexa. This indicates a parasite-specific function for these regulated genes, and the authors hypothesize that a divergent regulatory machinery might be responsible for the regulation of their expression.

In *C. parvum*, whole genome sequences revealed a similar paucity of known transcription factors encoded in the genome. Experimental evaluation of gene expression is most difficult in this parasite because of the absence of a continuous *in vitro* culture system. Preliminary analyses via RT-PCR experiments revealed that coordinately expressed genes did not necessarily occupy a clustered region in the genome, indicating the presence of gene-specific regulatory elements (Abrahamsen et al. 2004). We have examined the sequences upstream of genes that are functionally related in *C. parvum* and find a correlation between functionally related genes and the presence of conserved sequence motifs in the upstream regions (Mullapudi et al. 2007). Functionally related genes for some of the metabolic activities are co-expressed in sub-sets and unlike what was observed in *Plasmodium* and *Toxoplasma*, show biphasic expression, with their transcripts reaching peak expression levels at more than one time point in the post-infection life-cycle (Mullapudi et al. 2007), (Chapter 3).

2.1.2 Evidence for gene regulation at multiple levels in *P. falciparum* and *T. gondii*

Chromatin-modifying enzymes and epigenetic regulation:

The basic repeat unit of chromatin is the nucleosome, which consists of 146 bp of DNA wrapped around an octamer of core histone proteins: H2A, H2B, H3 and H4, each present in two copies. These histone proteins can be modified post-translationally in several ways such as methylation and acetylation of specific residues. Depending on the modification, they render the associated chromatin into a transcriptionally active or repressed state (Horn and Peterson 2002; Fischle et al. 2003). Both *P. falciparum* and *T. gondii* encode canonical forms of the core histone proteins (but seem to lack the linker histone H1) as well as specific histone variants (Sullivan 2003; Sullivan et al. 2003; Miao et al. 2006). They also contain a rich repertoire of histone-modifying enzymes including histone deacetylases, acetylases and other histone remodeling factors such as components of the Swi/Snf complex found in yeast (Bhatti et al. 2006; Sullivan and Hakimi 2006). Furthermore, it was found that histone deacetylase (HDAC) activity was necessary for the survival of the parasites, as drugs such as apicidin with anti-HDAC activity could inhibit growth of the parasites (Darkin-Rattray et al. 1996).

Recent studies in *P. falciparum* have shown that chromatin-mediated processes play an important role in the regulation of *var* gene expression, a gene family involved in antigenic variation. Antigenic variation in *P. falciparum* is brought about by three highly variable gene families termed *var*, *rif* and *stevors*, mostly located in clusters in sub-telomeric regions. The *var* gene family encodes proteins (PfEMPI- erythrocyte membrane protein I) that are exported to the surface of the infected red blood cell where they mediate adherence to host endothelial receptors and help in sequestration of infected cells (Kyes et al. 2007). Transcriptional switching between the *var* genes causes antigenic variation and helps the parasite evade the host immune system.

Exclusive expression of a single *var* gene is facilitated by a cooperative silencing mechanism that involves two *cis*-acting elements in the var gene promoter and first intron (Duraisingh et al. 2005). However, it was shown that effective silencing of the *var* gene can only be achieved when the DNA is associated with chromatin (Deitsch et al. 2001). More recently, in a series of chromatin-immunoprecipitation (chIP) experiments that investigated associated changes in the chromatin in the context of var gene silencing (Chookajorn et al. 2007) it was shown that an epigenetic mark (methylation of histone H3 at lysine K9) was associated with *var* gene silencing and contributed to "epigenetic memory" in the malarial parasite. This is one of the first such instances of epigenetic memory being reported to operate in a unicellular eukaryote, and the underlying mechanisms are yet to be elucidated.

In *T. gondii*, it has been shown that histone acetylation is linked to stage-differentiation (Saksouk et al. 2005). Through chIP, it was found that acetylation and methylation of *T. gondii* histones correlated with stage-specific gene expression in the parasite. Saksouk *et al* examined the promoters of stage-specific genes in the tachyzoite stage and established that tachyzoite-specific promoters were associated with TgGCN5 (acetylation) while bradyzoite specific promoters were associated with hypoacetylation and HDAC activity. They also detected histone 3 methylation in tachyzoite promoters and found that the inhibition of an arginine methyl transferase (TgCARM1) in *T.gondii* induced cyst formation. The extensive repertoire of histone modifying enzymes found in *T. gondii* (Sullivan and Hakimi 2006) combined with the apparent lack of specialized transcriptional machinery lends credence to the idea that epigenetic regulation is responsible for a considerable part of gene regulation.

Transcriptional regulation:

In order for a gene to be expressed, it is first transcribed by the RNA polymerase complex and associated proteins that are required for transcript initiation and elongation. Regulation of gene expression at the transcriptional level is brought about by the interplay between proteins (transcription factors) and recognition sites on the DNA (*cis*-regulatory elements) that recruit transcription factors (Eulgem 2001; Thomas and Chiang 2006). Typical eukaryotic transcription factors belong to two categories: the basal machinery consists of general transcription factors (GTFs) and transcriptional coactivators. GTFs bind to the core promoter of a gene and are required for baseline transcription of genes while transcriptional coactivators are typically recruited via other proteins and are necessary for activation of transcription. The second category of transcription factors are called specialized transcription factors that are recruited in a gene-specific manner to achieve preferential gene expression (Hampsey 1998).

A comparative analysis of the transcriptional machinery of different protozoan genomes (including *P. falciparum* and *T. gondii*) showed that they appear to possess a smaller number of conserved general transcription factors. This was deduced by sequence comparison to known eukaryotic transcription factors (Meissner and Soldati 2005). Hidden Markov Model (HMM) based searches of the *P. falciparum* genome using a representative set of eukaryotic transcription associated proteins (TAPs) revealed that a very small fraction of encoded proteins (71 in total) contained sequence similarity to similar proteins in other eukaryotes (Coulson et al. 2004). However, subsequent studies in *P. falciparum* that employed more sophisticated search methods based on conservation of secondary structure identified more TAPs than were originally thought to be present (Callebaut et al. 2005). In these studies, the authors employed Hydrophobic Cluster Analysis (HCA) coupled with profile-based searches to unearth divergent proteins in P.falciparum that could serve as components of the basal transcriptional machinery. The HCA

method involved analysis of *P. falciparum* proteins based on their secondary rather than primary structure. Conserved secondary structure is more indicative of function than primary structure (Chothia and Lesk 1986). This method gets around the problems posed by the strong AT-bias in the *P. falciparum* genome and the potentially high divergence of apicomplexan transcription factors and thus was able to attribute transcription-related function to several orphan proteins in the *P. falciparum* genome. From all these studies, it can be seen that most of the general transcription factors (GTFs) have been identified in *T. gondii* and *P. falciparum*, while transcriptional coactivators do not seem to be present to the same extent (See Table 2.1). For example, several TBP associated factors (TAFs) that are essential for transcriptional activation in metazoans seem to be missing in the Apicomplexa. Specialized transcription factors have also not been identified in the apicomplexan genomes. Whole genome sequences for the apicomplexans *P. falciparum* and *C. parvum* have revealed a glaring paucity of specialized transcription factors encoded in the genome (Gardner et al. 2002; Abrahamsen et al. 2004). One specialized transcription factor PfMyb1 has been identified cloned and characterized in *P. falciparum* (Boschet et al. 2004; Gissot et al. 2005). The PfMyb1 protein is similar in sequence to the Myb family of transcription factors conserved in all eukaryotes and was found to bind specifically to a canonical Myb regulatory element (MRE) in the chicken gene promoter, as well as to putative MREs annotated within *Plasmodium* promoters. This is the only specialized transcription factor to date that has been experimentally characterized in the Apicomplexa. Besides PfMyb1, A study that used highly sensitive profile-based searches led to the discovery of a lineage specific family of DNA-binding proteins, the ApiAP2 family which show resemblance to a family of transcription factors in plants (AP2 – Apetala2-integrase binding

domain) and are likely to function as specific transcription factors in the Apicomplexa (Balaji et al. 2005).

At the level of promoter organization and conserved binding sites for transcription factors, independent gene-specific studies in *T.gondii* as well as *P. falciparum* showed the absence of canonical eukaryotic *cis*-regulatory elements in promoter sequences (Mercier et al. 1996; Militello et al. 2004). In *P. falciparum*, a few gene-specific *cis*-acting elements have been identified in individual genes that are expressed in a stage-specific manner. One of the earliest such reports was on the *GBP130* (glycophorin binding protein 130): a gene known to be developmentally regulated at the level of transcription initiation. Using traditional promoter deletion approaches, a 300 bp repeat element present twice in its promoter region was found to be significant in downstream gene expression. Deletion of one of these repeat elements (the 5′ element) had a positive influence on gene expression, whereas deletion of both elements had a negative influence on expression. Further mutagenesis analyses identified a 5 bp sequence 5′GTATT that was essential for promoter activity and specifically bound to protein from *P. falciparum* nuclear extracts (Lanzer et al. 1992; Lanzer et al. 1993; Horrocks and Lanzer 1999).

Analysis of two genes *pfs16* and *pfs25* that are markers of sexual development in *P. falciparum* show the involvement of upstream DNA sequences in the specific expression of these genes. An 8 bp motif 5′AAGGAATA was identified by promoter deletion experiments and found to be necessary for the enhanced transcription of *pfs25*. Subsequent gel-shift assays also demonstrated the binding of a mosquito stage-specific protein PAF1 to this 8 bp motif. It is interesting to note that other eukaryotic motifs identified in the promoter of pfs25 did not seem to have an effect on gene expression (Dechering et al. 1997).

Similarly, conserved sequence elements in the promoter and the first intron of the *var* genes were found to be important for the expression and regulation of the *var* gene family involved in antigenic variation. These elements were also shown to bind specifically to proteins from parasite nuclear extract (Vazquez-Macias et al. 2002; Calderwood et al. 2003; Voss et al. 2006). In the case of *var* gene expression, recent studies have indicated that epigenetic regulation may also have a significant role to play in the allelic exclusion and silencing of the *var* genes (Chookajorn et al. 2007) as discussed earlier.

Other *cis*-acting elements important for transcriptional regulation include a 20 bp element showing dual symmetry found upstream of the KAHRP gene in *P. falciparum*, another developmentally regulated gene (Lanzer et al. 1992) and a 24 bp element found to be necessary for specific binding in the promoter region of CDP-diacyl glycerol synthase (Osta et al. 2002). Bioinformatics analyses of *P. falciparum hsp* genes led to the discovery of a G-box motif 5′(A/G)NGGGG(C/A) conserved among 18 *hsp* genes. This motif was also found in the promoter regions of corresponding orthologs in *P. yoelii*, *P. berghei* and *P. vivax*. The G-box was present in two copies in a tandem palindromic organization, and reporter assays using *hsp*86 promoter with and without the G-box indicated a definite role for the G-box in downstream gene expression (Militello et al. 2004)

In *T. gondii*, the earliest promoter-deletion studies identified a 7 bp motif 5′(A/T)GAGACG within a repeating 27 bp unit in the promoters of the SAG gene family (Soldati and Boothroyd 1995). This motif was found to behave as a selector of transcriptional initiation as well as an enhancer. Subsequent studies in other genes also identified the same element in promoter regions, showing an effect on downstream gene expression in reporter assays (Mercier et al. 1996). Recently, the promoter of the SAG1 gene was analyzed by deletion

mutagenesis and electrophoresis mobility shift assay (EMSA) (Zhang et al. 2007). These studies established the importance of a 10 bp element 5′TGGGCAGAGC (different from the 7 bp motif reported earlier) in both promoter activity and in protein-binding as was seen in EMSA with tachyzoite nuclear extract.

Other gene-specific studies reported the presence of different conserved *cis*-acting elements that were important for downstream gene expression as evidenced in reporter assays. These include a GC-rich palindromic region 5′CGCAGCG identified upstream of the BAG-1 gene (Bohne et al. 1997), and a 7 bp Inr like element 5′TCAGTTT identified upstream of the nucleoside triphosphate hydrolase gene (Nakaar et al. 1998).

In a study of the *T. gondii* hsp70 promoter by deletion analyses, a series of GAA repeats similar to the known eukaryotic heat shock element (HSE) were found (Ma et al. 2004). An Sp-1 like element 5′CCGGGG was also reported in this promoter region, and these regions play a significant role in parasite response to pH stress. In subsequent gel-shift assays, Ma *et al* identified a specific sequence in the *hsp70* promoter that specifically bound to protein from *T. gondii* bradyzoite nuclear extracts. This sequence is not a classical HSE and a survey of the *T. gondii* genome did not reveal the presence of classical heat shock proteins that might bind it.

Toxoplasma gondii expresses two distinct isoforms of the glycolytic enzyme enolase (*ENO1* and *ENO2*) depending upon the developmental stage. By semi-quantitative RT-PCR, it was found that *ENO1* mRNA was exclusively found in bradyzoites whereas *ENO2* mRNA was only found in tachyzoites (Dzierszinski et al. 1999; Dzierszinski et al. 2001), indicating regulation at the transcriptional level. The two genes are found in the same locus in the *T. gondii* genome, positioned in a tandem array and share 75% amino acid sequence similarity of the coding regions. Examination of their promoter sequences and sequential deletion using reporter

constructs showed that their promoters contain specific sequence elements which can either positively or negatively affect downstream gene expression (Kibe et al. 2005). By using nuclear extracts from tachyzoites and bradyzoites, Kibe *et al* were able to map regions in these promoters that specifically bound to proteins. They found that a eukaryotic stress response element (STRE) in the *ENO1* promoter 5′ACAGGGGGA was responsible for the specific binding of nuclear extract from bradyzoites. Mutagenesis of this motif abolished binding. This STRE has also been found in the promoter of *hsp70* (Ma et al. 2004)

Ribosomal protein genes were found to be developmentally regulated in the coccidian parasites *T. gondii* and *Eimeria tenella* (Schaap et al. 2005). Two conserved *cis*-regulatory elements have been found to be over-represented in the promoters of all 79 cytoplasmic ribosomal protein-encoding genes in *T. gondii* (NF et al. 2006). These elements were detected by computational analyses but no biological significance for the motifs was reported.

Post-transcriptional regulation

Regulation of gene expression can take place at the post-transcriptional level through alternative splicing or by selective stabilization/degradation of messages (Day and Tuite 1998). In *P. falciparum*, various reports point to the significance of post-transcriptional regulation in the parasite's life-cycle. Recent microarray and proteomic analyses indicate that there is a delay between transcript synthesis and protein production, indicating the possible role of post-transcriptional regulatory methods in gene expression (Le Roch et al. 2003; Le Roch et al. 2004). There are at least two examples of pseudogenes being transcribed in *P. falciparum*. A homolog of the EBA175 protein, an erythrocyte binding protein was discovered and found to be transcribed but not translated in the intra-erythrocytic stage of the parasite (Triglia et al. 2001). It was discovered that this transcript contained several frame-shift mutations that indicate that the

gene was a pseudogene and not necessary for parasite growth and survival. Another example of a putative pseudogene which was transcribed in its entirety but not translated is the *Plasmodium* rhoptry protein *PfRH3* (Taylor et al. 2001). The retention and transcription of these pseudogenes poses a question about their functional relevance. Transcription of pseudogenes has been reported to have a regulatory role in mice (Yano et al. 2004), where it was found that pseudogenes generated anti-sense transcripts that controlled the expression of the sense transcript. It would be interesting to see if these pseudogenes in *P. falciparum* produce anti-sense transcripts with a similar regulatory role. SAGE analyses have revealed the prevalence of anti-sense transcripts in the *P. falciparum* genome on a genome-wide level (Gunasekera et al. 2004). 12% of the 17,000 SAGE tags (representing a total of 1500 ORFs) were found to represent anti-sense transcripts, and did not exhibit any defined pattern with respect to the genomic location of the genes. All anti-sense transcripts were derived from nuclear-encoded genes, and many of them were for proteolysis and translation-related genes in the parasite. An inverse relationship was observed between sense and anti-sense tags in terms of abundance, so that loci with a very small number of sense tags exhibited a large number of anti-sense tags. This led the authors to hypothesize that the anti-sense transcripts function in regulating gene expression by limiting the number of sense-transcripts produced from a particular locus. They propose mechanisms by which this could take place, such as binding and targeting the sense partner for degradation, interfering with transcriptional initiation and elongation and /or inhibiting translation of the sense-strand counterpart. As of now, no experimental evidence has been obtained to support these hypotheses.

Plasmodium falciparum is also found to encode members of the Puf RNA-binding protein family that bind to specific regions in the 3′ UTR of mRNAs and regulate message

stability and translation (Cui et al. 2002). Translational repression has also been shown to be essential for sexual development of the malarial parasite *Plasmodium bergheii* (Mair et al. 2006). Mair and coworkers identified an RNA helicase DOZ1 that is highly upregulated in female gametocytes and through knock-out studies showed that it played a central role in the silencing and maintenance of a steady-state level of gametocyte-specific transcripts. Finally, in a computational analysis of promoter regions of gene clusters found to be co-regulated in chloroquine-treated parasites, two outstanding motifs: 5′GAGAGAA and 5′ACTATAAAGA were identified that selectively formed RNA-protein complexes but not DNA-protein complexes, as evidenced by gel-shift assays (Gunasekera et al. 2007). These motifs were also found to affect gene expression in heterologous reporter assays. These data suggest a role for these motifs in gene expression, possibly at the mRNA level by interacting with specific RNA-binding proteins.

Comparatively lesser information is available on post-transcriptional methods of regulation in *T. gondii*. SAGE analyses did not report the presence of anti-sense transcripts in *T.gondii* as seen in *P. falciparum* (Radke et al. 2005). Two stage-specific genes have been found that are probably post-transcriptionally regulated at the mRNA level. The surface antigen P36 is a bradyzoite-specific protein. However, mRNA produced by the coding gene *BSR4* is found to be equally abundant in both stages, tachyzoite and bradyzoite, as evidenced by RT-PCR studies (Boothroyd et al. 1997). In the case of another bradyzoite-specific gene *LDH*, mRNA evidence for this isoenzyme was found in both tachyzoites and bradyzoites, although the protein product is exclusive to bradyzoites (Yang and Parmley 1995). These studies indicate that at least some genes are controlled at the level of post-transcription, possibly via interactions with UTR sequences or other methods.

2.1.3 Summary

Experimental evidence exists for gene regulation at various levels in the apicomplexan parasites *P. falciparum* and *T. gondii* as has been detailed from the existing literature. Large-scale expression analyses point to a clearly regulated transcriptome in a stage and time-dependent manner and individual gene-specific analyses indicate different levels of gene regulation. In the context of transcriptional regulation, whole genome sequence surveys reveal an apparent paucity of specialized transcription factors when compared to known eukaryotic transcription factors. Most of the components of the general transcription machinery have been identified by sequence-similarity based searches or by profile-based and secondary-structure based searches. Thus far, only one specific transcription factor has been identified in *P. falciparum* that has had its function validated experimentally. The idea that this phylum of parasites possesses highly divergent transcriptional machinery is gaining strength from the observations thus far. The work described in this thesis focuses on identifying novel *cis*-regulatory elements in the genomes of *C. parvum* and *T. gondii* and investigating their role in gene regulation to help us to better understand the nature of gene regulation in the Apicomplexa.

2.2 An introduction to computational approaches to *cis*-regulatory element detection

Given the absence of known *cis*-regulatory elements in apicomplexan parasites and the apparent scarcity of known transcription factors, genome sequence provides us with a starting point to investigate *cis*-element-mediated regulation by identifying conserved upstream elements that could serve as binding sites for putative transcription factors in these parasites. This section describes some of the commonly used bioinformatics methods developed to detect *cis*-regulatory elements and their underlying principles.

2.2.1 Definition of a *cis*-regulatory element

Cis-regulatory elements are defined as short 5-20 bp regions in the DNA that commonly act as transcription factor binding sites (Fickett and Hatzigeorgiou 1997) Transcription factors recognize these sites in the DNA and bind to them preferentially, to exert their effect on the expression of the coding sequence associated with the regulatory region. These effects can be either positive (enhancement of transcription) or negative (transcriptional repression).

2.2.2 *Cis*-regulatory elements are "fuzzy" entities

It has been observed that regulatory regions are very well conserved between some species, implying an important biological role for these sequences (Lee et al. 2005). This has especially been observed in metazoan genomes, where blocks of conserved non-coding sequences have been found between rodent and human genes, and these regions were found to contain regulatory information (Hardison et al. 1997). This is also seen in inter-species comparison of yeast species (Doniger et al. 2005), allowing for the establishment of consensus binding sites based on conserved sequences of *cis*-elements of orthologous genes (Kellis et al. 2003). However, the sequences of regulatory regions do not have to be 100% conserved and in practice they often show a considerable degree of variability (Figure 2.1). This variation arises due to two reasons. Firstly, non-coding regions are not under the same evolutionary constraints as coding regions (Dermitzakis and Clark 2002). Changes in their sequence can be tolerated more easily due to compensatory changes elsewhere in the genome (Ludwig et al. 2000). Such compensatory changes are facilitated in part by the small size of individual regulatory elements. This makes it possible for a new site to arise in the region by chance that can functionally replace the original element. Alternatively, a mutation in the corresponding transcription factor which allows it to recognize an altered binding site can eventually lead to the selection of the new binding site. On the other hand, variations in significant positions of binding site sequence can in

fact be used to fine tune the protein-DNA interaction, leading to varying magnitudes of promoter activity (Natoli 2004) . Secondly, the biological function of a regulatory motif (to behave as a binding site for a protein) renders it some amount of flexibility at the sequence level. Protein-DNA interactions do not have to adhere to a strict sequence-based code (Benos et al. 2002). Not all the positions in a given binding site contribute to the same extent to the binding of its cognate trancription factor. The efficiency of a protein binding to a specific DNA sequence is dictated by higher order constraints such as the mode of occupation of the site (dimerization, protein-protein interactions, etc.) and the physical state of chromatin and DNA in the vicinity of the promoter. Binding is also governed by thermodynamic properties. Chemical complementarity between the protein and DNA dictates the affinity of the protein for the DNA- these forces can also exist in non-specific interactions. Hydrogen bonding, on the other hand, plays an important role in sequence-specific recognition and dictates the specificity of such interactions (Benos et al. 2002). All of these properties obviate the need for a strict sequence-level conservation of the binding site and make a *cis*-regulatory element a "fuzzy" entity, which does not adhere to a precise set of rules. Consequently, the computational detection of such elements remains a non-trivial problem and no single approach can provide the complete or correct solution as yet (Ohler and Niemann 2001; Vavouri and Elgar 2005).

2.2.3 Approaches to computational detection of *cis*-regulatory elements

Sequence-based approaches to detect *cis*-regulatory elements need to go beyond simple local alignment approaches such as BLAST that can detect local regions of similarity in aligned sequences. This is because the relative position of *cis*-regulatory elements is not always conserved (See Figure 2-2). Additionally their small size and sequence degeneracy makes them unsuitable for detection via alignment methods. Several algorithms have been developed for the

detection of patterns in sequence data either by *de novo* methods or by making use of *a priori* knowledge of the nature of the pattern. With the increased availability of whole genome sequences, these algorithms have become very popular for the identification of regulatory elements within non-coding sequences and have met with varying degrees of success (Tompa et al. 2005).

De novo pattern finding: looking for a needle in a haystack

When no prior information about a putative *cis*-regulatory element is available, pattern-finding programs are employed to take a blind approach and identify over-represented, statistically significant sequence motifs from a group of provided sequences, in this case upstream promoter sequences. Different programs use different algorithms to detect such putative regulatory elements. Two such programs used in this study are MEME (Multiple Expectation Maximization for Elicitation of Motifs) (Bailey and Elkan 1994) and AlignACE (Aligns Nucleic Acids Conserved Elements) (Hughes et al. 2000). These programs build either deterministic (MEME) or probabilistic (AlignACE) models of the sequence motif and then evaluate the input data for the best solutions or best fit, operating on certain input parameters such as length of the motif, occurrence, etc. MEME scans the input sequences, assigning sites to a motif and builds on it as more occurrences of the same motif are found (D'Haeseleer 2006). The final model is then evaluated and reported along with its scores. AlignACE uses a probabilistic method called Gibbs sampling to report motifs that are over-represented in the input set over random. Other programs like TEIRESIAS (Rigoutsos and Floratos 1998) use enumerative approaches, using a dictionary based method to count all occurrences of n-mers in the given search space, and reporting those that are over-represented. In all of these cases, several motifs can be reported, and the problem of sifting "real" motifs from noise is a

challenging one, that can be addressed to some extent by evaluating other criteria that contribute to the "quality" of a motif as described below.

Measuring motif quality

Motifs are typically reported by programs in order of their statistical significance, which can be computed in different ways. The two most common properties used towards this calculation are the information content of the motif and it's degree of over-representation compared to the background. Information content is a measure of the overall frequencies of a base at a given position compared to the overall frequency of the base in the group of input sequences. Over-representation of a motif is measured in comparison to a background set which is used as the "null set" against which nucleotide frequencies are measured. This background set can be specified to most pattern finding programs. It can be a set of sequences constituting the whole genome, intergenic regions from the whole genome or any other set of sequences as desired. While providing for a robust way to measure over-representation, this method may not always give the best results. In case of AT-rich genomes such as *P. falciparum* and *C. parvum*, any AT-rich motif will be regarded as random by the program due to the AT-richness of the background, thus skewing the output to report GC-rich motifs. Secondly, biological motifs are not necessarily over-represented in the genome, and it is very possible that a given motif may be functional in several different contexts and environments, irrespective of the surrounding sequence. Hence, it becomes imperative that informative rules are established to narrow down candidate motifs from a given output. These additional rules can take into account several other properties such as positional information relative to transcriptional or translational start, evolutionary conservation from related species, enrichment of motifs at specific sites near co-regulated genes, combinatorial occurrences of motifs and clustering of motifs in regulatory

regions to help identify significant candidates. In chapters 3 and 4, we discuss the establishment of such rules for the purpose of detecting biologically significant motifs.

2.2.4 Augmentation of computational methods using biological insight

De novo approaches are the methods of choice when nothing is known about the nature or pattern of *cis*-regulatory elements in a genome. However, these methods can be improved by adding biological information of various kinds. Sequence information from a closely related species can help us to identify evolutionarily conserved regions of DNA that might indicate biological function (Zhang and Gerstein 2003). The choice of species for such comparative analysis (phylogenetic footprinting) is important, since species that are too closely related will not provide any additional information, whereas species that are too distant may not enhance the sample space in an informative way Expression information about the groups of genes under study can be used to cluster the genes into co-regulated groups to help find motifs associated with a particular pattern of expression. Finally, in the absence of any such information, the results of computational pattern-finding provide us with starting points to examine the biological role of the motif in experiments such as reporter-assays and gel-shift assays (Mullapudi et al, 2006).

2.3 Summary

The regulatory content of non-coding DNA of the genome is known to shape genome architecture (Nelson et al. 2004) In mammals, it has been observed that genes involved in complex adaptive processes have highly conserved upstream regions. In compact genomes such as those of the Apicomplexa, conserved non-coding regions could be purported to have some kind of biological significance such as regulatory function, which would explain their retention in spite of genome compaction. To detect such regulatory elements without any *a priori*

knowledge, several computational tools are available to detect over-represented conserved DNA motifs that can be assigned a putative regulatory function. It is important to tweak and optimize user-defined parameters for each algorithm, so that one can examine all possible solutions to identify biologically significant candidates. Finally, *in silico* methods while providing one with a starting point to explore regulatory elements in a genome, have to be supplemented by biological insight and experimental confirmation to better improve their applicability and predictive power.

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G et al. (2004) Complete Genome Sequence of the Apicomplexan, *Cryptosporidium parvum*. Science (New York, NY 304: 441-445.

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.

Balaji S, Babu MM, Iyer LM, Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic acids research 33(13): 3994-4006.

Benos PV, Lapedes AS, Stormo GD (2002) Is there a code for protein-DNA recognition? Probab(ilistical)ly. Bioessays 24(5): 466-475.

Bhatti MM, Livingston M, Mullapudi N, Sullivan WJ, Jr. (2006) Pair of unusual GCN5 histone acetyltransferases and ADA2 homologues in the protozoan parasite *Toxoplasma gondii*. Eukaryotic cell 5(1): 62-76.

Bohne W, Wirsing A, Gross U (1997) Bradyzoite-specific gene expression in *Toxoplasma gondii* requires minimal genomic elements. Mol Biochem Parasitol 85(1): 89-98.

Boothroyd JC, Black M, Bonnefoy S, Hehl A, Knoll LJ et al. (1997) Genetic and biochemical analysis of development in *Toxoplasma gondii*. Philos Trans R Soc Lond B Biol Sci 352(1359): 1347-1354.

Boschet C, Gissot M, Briquet S, Hamid Z, Claudel-Renard C et al. (2004) Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 *Plasmodium falciparum* clones. Mol Biochem Parasitol 138(1): 159-163.

Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J et al. (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. PLoS Biol 1(1): 5.

Calderwood MS, Gannoun-Zaki L, Wellems TE, Deitsch KW (2003) *Plasmodium falciparum* var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron. The Journal of biological chemistry 278(36): 34125-34132.

Callebaut I, Prat K, Meurice E, Mornon JP, Tomavo S (2005) Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes. BMC genomics 6: 100.

Chookajorn T, Dzikowski R, Frank M, Li F, Jiwani AZ et al. (2007) Epigenetic memory at malaria virulence genes. Proceedings of the National Academy of Sciences of the United States of America 104(3): 899-902.

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. The EMBO journal 5(4): 823-826.

Cleary MD, Singh U, Blader IJ, Brewer JL, Boothroyd JC (2002) *Toxoplasma gondii* asexual development: identification of developmentally regulated genes and distinct patterns of gene expression. Eukaryotic cell 1(3): 329-340.

Coulson RM, Hall N, Ouzounis CA (2004) Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. Genome research 14(8): 1548-1554.

Cui L, Fan Q, Li J (2002) The malaria parasite *Plasmodium falciparum* encodes members of the Puf RNA-binding protein family with conserved RNA binding activity. Nucleic acids research 30(21): 4607-4617.

D'Haeseleer P (2006) What are DNA sequence motifs? Nature biotechnology 24(4): 423-425.

Darkin-Rattray SJ, Gurnett AM, Myers RW, Dulski PM, Crumley TM et al. (1996) Apicidin: a novel antiprotozoal agent that inhibits parasite histone deacetylase. Proceedings of the National Academy of Sciences of the United States of America 93(23): 13143-13147.

Day DA, Tuite MF (1998) Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. J Endocrinol 157(3): 361-371.

Dechering KJ, Thompson J, Dodemont HJ, Eling W, Konings RN (1997) Developmentally regulated expression of pfs16, a marker for sexual differentiation of the human malaria parasite *Plasmodium falciparum*. Mol Biochem Parasitol 89(2): 235-244.

Deitsch KW, Calderwood MS, Wellems TE (2001) Malaria. Cooperative silencing elements in var genes. Nature 412(6850): 875-876.

Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Molecular biology and evolution 19(7): 1114-1121.

Doniger SW, Huh J, Fay JC (2005) Identification of functional transcription factor binding sites using closely related *Saccharomyces* species. Genome research 15(5): 701-709.

Duraisingh MT, Voss TS, Marty AJ, Duffy MF, Good RT et al. (2005) Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum*. Cell 121(1): 13-24.

Dzierszinski F, Mortuaire M, Dendouga N, Popescu O, Tomavo S (2001) Differential expression of two plant-like enolases with distinct enzymatic and antigenic properties during stage conversion of the protozoan parasite *Toxoplasma gondii*. Journal of molecular biology 309(5): 1017-1027.

Dzierszinski F, Popescu O, Toursel C, Slomianny C, Yahiaoui B et al. (1999) The protozoan parasite *Toxoplasma gondii* expresses two functional plant-like glycolytic enzymes. Implications for evolutionary origin of apicomplexans. The Journal of biological chemistry 274(35): 24888-24895.

Eulgem T (2001) Eukaryotic transcription factors. Genome Biology 2(2): reports0004.

Fickett JW, Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. Genome research 7(9): 861-878.

Fischle W, Wang Y, Allis CD (2003) Histone and chromatin cross-talk. Current opinion in cell biology 15(2): 172-183.

Gardner MJ, Hall N, Fung E, White O, Berriman M et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419(6906): 498-511.

Gissot M, Briquet S, Refour P, Boschet C, Vaquero C (2005) PfMyb1, a *Plasmodium falciparum* transcription factor, is required for intra-erythrocytic growth and controls key genes for cell cycle regulation. Journal of molecular biology 346(1): 29-42.

Gunasekera AM, Patankar S, Schug J, Eisen G, Kissinger J et al. (2004) Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. Mol Biochem Parasitol 136(1): 35-42.

Gunasekera AM, Myrick A, Militello KT, Sims JS, Dong CK et al. (2007) Regulatory motifs uncovered among gene expression clusters in *Plasmodium falciparum*. Mol Biochem Parasitol.

Hampsey M (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol Mol Biol Rev 62(2): 465-503.

Hardison RC, Oeltjen J, Miller W (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome research 7(10): 959-966.

Horn PJ, Peterson CL (2002) Molecular biology. Chromatin higher order folding--wrapping up transcription. Science (New York, NY 297(5588): 1824-1827.

Horrocks P, Lanzer M (1999) Mutational analysis identifies a five base pair *cis*-acting sequence essential for GBP130 promoter activity in *Plasmodium falciparum*. Mol Biochem Parasitol 99(1): 77-87.

Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. Journal of molecular biology 296(5): 1205-1214.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423(6937): 241-254.

Kibe MK, Coppin A, Dendouga N, Oria G, Meurice E et al. (2005) Transcriptional regulation of two stage-specifically expressed genes in the protozoan parasite *Toxoplasma gondii*. Nucleic acids research 33(5): 1722-1736.

Kyes S, Christodoulou Z, Pinches R, Kriek N, Horrocks P et al. (2007) *Plasmodium falciparum* var gene expression is developmentally controlled at the level of RNA polymerase II-mediated transcription initiation. Mol Microbiol 63(4): 1237-1247.

Lanzer M, de Bruin D, Ravetch JV (1992) A sequence element associated with the *Plasmodium falciparum* KAHRP gene is the site of developmentally regulated protein-DNA interactions. Nucleic acids research 20(12): 3051-3056.

Lanzer M, Wertheimer SP, de Bruin D, Ravetch JV (1993) *Plasmodium*: control of gene expression in malaria parasites. Experimental parasitology 77(1): 121-128.

Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK et al. (2003) Discovery of Gene Function by Expression Profiling of the Malaria Parasite Life-cycle. Science (New York, NY.

Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A et al. (2004) Global analysis of transcript and protein levels across the *Plasmodium falciparum* life-cycle. Genome Res 14(11): 2308-2318.

Lee S, Kohane I, Kasif S (2005) Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. BMC genomics 6: 168.

Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403(6769): 564-567.

Ma YF, Zhang Y, Kim K, Weiss LM (2004) Identification and characterisation of a regulatory region in the *Toxoplasma gondii* hsp70 genomic locus. International journal for parasitology 34(3): 333-346.

Mair GR, Braks JA, Garver LS, Wiegant JC, Hall N et al. (2006) Regulation of sexual development of *Plasmodium* by translational repression. Science (New York, NY 313(5787): 667-669.

Manger ID, Hehl A, Parmley S, Sibley LD, Marra M et al. (1998) Expressed sequence tag analysis of the bradyzoite stage of *Toxoplasma gondii*: identification of developmentally regulated genes. Infection and immunity 66(4): 1632-1637.

Meissner M, Soldati D (2005) The transcription machinery and the molecular toolbox to control gene expression in *Toxoplasma gondii* and other protozoan parasites. Microbes Infect 7(13): 1376-1384.

Mercier C, Lefebvre-Van Hende S, Garber GE, Lecordier L, Capron A et al. (1996) Common *cis*-acting elements critical for the expression of several genes of *Toxoplasma gondii*. Mol Microbiol 21(2): 421-428.

Miao J, Fan Q, Cui L, Li J, Li J et al. (2006) The malaria parasite *Plasmodium falciparum* histones: organization, expression, and acetylation. Gene 369: 53-65.

Militello KT, Dodge M, Bethke L, Wirth DF (2004) Identification of regulatory elements in the *Plasmodium falciparum* genome. Mol Biochem Parasitol 134(1): 75-88.

Mullapudi N, Lancto CA, Abrahamsen MS, Kissinger JC (2007) Identification of putative *cis*-regulatory elements in *Cryptosporidium parvum* by de novo pattern finding. BMC genomics 8: 13.

Nakaar V, Bermudes D, Peck KR, Joiner KA (1998) Upstream elements required for expression of nucleoside triphosphate hydrolase genes of *Toxoplasma gondii*. Mol Biochem Parasitol 92(2): 229-239.

Natoli G (2004) Little things that count in transcriptional regulation. Cell 118(4): 406-408.

Nelson CE, Hersh BM, Carroll SB (2004) The regulatory content of intergenic DNA shapes genome architecture. Genome Biol 5(4): R25.

NF VANP, Welagen J, Vermeulen AN, Schaap D (2006) The complete set of *Toxoplasma gondii* ribosomal protein genes contains two conserved promoter elements. Parasitology 133(Pt 1): 19-31.

Ohler U, Niemann H (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. Trends Genet 17(2): 56-60.

Osta M, Gannoun-Zaki L, Bonnefoy S, Roy C, Vial HJ (2002) A 24 bp *cis*-acting element essential for the transcriptional activity of *Plasmodium falciparum* CDP-diacylglycerol synthase gene promoter. Mol Biochem Parasitol 121(1): 87-98.

Perkins FO, Barta JR, Clopton RE, Peirce MA, Upton SJ (2000) Phylum Apicomplexa. In: Protozoologists So, editor. The Illustrated Guide to the Protozoa. second ed. Lawrence: Allen Press Inc. pp. 190-369.

Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS et al. (2005) The transcriptome of *Toxoplasma gondii*. BMC Biol 3: 26.

Rigoutsos I, Floratos A (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics (Oxford, England) 14(1): 55-67.

Saksouk N, Bhatti MM, Kieffer S, Smith AT, Musset K et al. (2005) Histone-modifying complexes regulate gene expression pertinent to the differentiation of the protozoan parasite *Toxoplasma gondii*. Molecular and cellular biology 25(23): 10301-10314.

Schaap D, Arts G, van Poppel NF, Vermeulen AN (2005) De novo ribosome biosynthesis is transcriptionally regulated in Eimeria tenella, dependent on its life-cycle stage. Mol Biochem Parasitol 139(2): 239-248.

Soldati D, Boothroyd JC (1995) A selector of transcription initiation in the protozoan parasite *Toxoplasma gondii*. Molecular and cellular biology 15(1): 87-93.

Sullivan WJ, Jr. (2003) Histone H3 and H3.3 variants in the protozoan pathogens *Plasmodium falciparum* and *Toxoplasma gondii*. DNA Seq 14(3): 227-231.

Sullivan WJ, Jr., Hakimi MA (2006) Histone mediated gene activation in *Toxoplasma gondii*. Mol Biochem Parasitol 148(2): 109-116.

Sullivan WJ, Jr., Monroy MA, Bohne W, Nallani KC, Chrivia J et al. (2003) Molecular cloning and characterization of an SRCAP chromatin remodeling homologue in *Toxoplasma gondii*. Parasitology research 90(1): 1-8.

Taylor HM, Triglia T, Thompson J, Sajid M, Fowler R et al. (2001) *Plasmodium falciparum* homologue of the genes for *Plasmodium* vivax and *Plasmodium yoelii* adhesive proteins, which is transcribed but not translated. Infection and immunity 69(6): 3635-3645.

Thomas MC, Chiang CM (2006) The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol 41(3): 105-178.

Tompa M, Li N, Bailey TL, Church GM, De Moor B et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nature biotechnology 23(1): 137-144.

Triglia T, Thompson J, Caruana SR, Delorenzi M, Speed T et al. (2001) Identification of proteins from *Plasmodium falciparum* that are homologous to reticulocyte binding proteins in *Plasmodium* vivax. Infection and immunity 69(2): 1084-1092.

van Driel R, Fransz PF, Verschure PJ (2003) The eukaryotic genome: a system regulated at different hierarchical levels. Journal of cell science 116(Pt 20): 4067-4075.

van Noort V, Huynen MA (2006) Combinatorial gene regulation in *Plasmodium falciparum*. Trends Genet 22(2): 73-78.

Vavouri T, Elgar G (2005) Prediction of *cis*-regulatory elements using binding site matrices--the successes, the failures and the reasons for both. Current opinion in genetics & development 15(4): 395-402.

Vazquez-Macias A, Martinez-Cruz P, Castaneda-Patlan MC, Scheidig C, Gysin J et al. (2002) A distinct 5' flanking var gene region regulates *Plasmodium falciparum* variant erythrocyte surface antigen expression in placental malaria. Mol Microbiol 45(1): 155-167.

Voss TS, Healer J, Marty AJ, Duffy MF, Thompson JK et al. (2006) A var gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria. Nature 439(7079): 1004-1008.

Yang S, Parmley SF (1995) A bradyzoite stage-specifically expressed gene of *Toxoplasma gondii* encodes a polypeptide homologous to lactate dehydrogenase. Mol Biochem Parasitol 73(1-2): 291-294.

Yano Y, Saito R, Yoshida N, Yoshiki A, Wynshaw-Boris A et al. (2004) A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. Journal of molecular medicine (Berlin, Germany) 82(7): 414-422.

Zhang J, Gu Q, Hou X, Zhou H, Cong H et al. (2007) Identification of a necessary element for *Toxoplasma gondii* SAG1 gene expression. Experimental parasitology.

Zhang Z, Gerstein M (2003) Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. Journal of biology 2(2): 11.

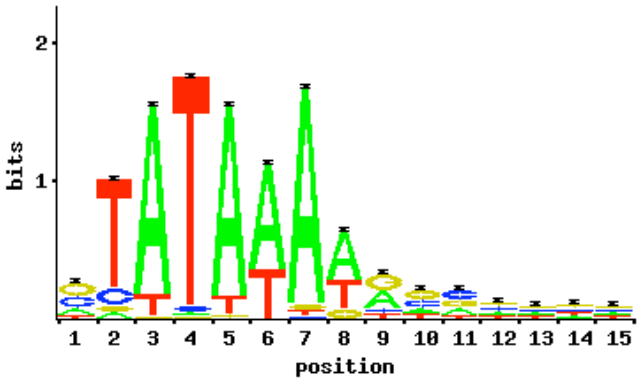
Table 2.1 Components of the basal transcriptional machinery identified in the Apicomplexa

A list of the basal transcriptional machinery in eukaryotes along with their sub-units if any (names of the proteins in *Saccharomyces cerevisiae* have been used) and the presence or absence of these counterparts in the Apicomplexans as detected by BLAST searches (Column I) (Meissner and Soldati, 2005) or as detected by profile searches accompanied by HCA (Column II) (Callebaut et al, 2005). "+" indicates that the counterpart was detected in either *T. gondii* or *P. falciparum*, "–" indicates that a significant match could not be found in either *P. falciparum* or *T. gondii*.

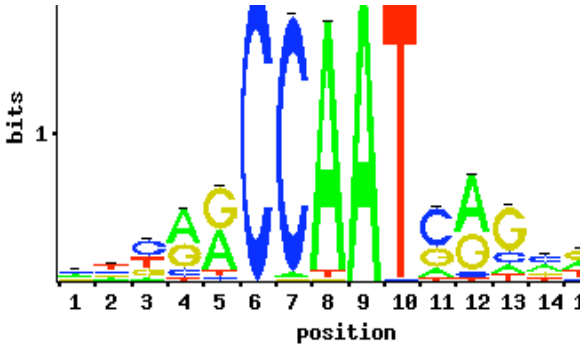
Table 2.1 Components of the basal transcriptional machinery identified in the Apicomplexa

Eukaryotic element	Sub-units	Apicomplexa	
General Transcriptional Factors		I	II
RNA Polymerase II holoenzyme	(12 subunits)	+	
TBP		+	
TFIIB		+	
TFIIE	TFA1	+	
	TFA2 b	–	+
TFIIF	**TFG1**	–	
	TFG2 b	–	+
	TFG3	+	
TFIIH	TFB1	–	+
	TFB2	+	
	TFB3	–	
	TFB4	+	
	TFB5	–	+
	RAD3	+	
	SSL1	+	
	SSL2	+	
	KIN28	+	
	CCL1	+	
Transcriptional Coactivators			
TFIIA	a	–	+
	b	–	
	g	–	+
TFIID	TAF145	–	+
	TAF150	–	+
	TAF130	+	
	TAF90	–	
	TAF67	+	
	TAF61	–	
	TAF60	+	
	TAF47	–	
	TAF40	–	
	TAF30	–	
	TAF25	–	+
	TAF19	–	
	TAF17	–	
SAGA Complex	GCN5	+	
	ADA1	–	
	ADA2	+	
	ADA3	–	

Figure 2.1 Sequence degeneracy in known binding sites: Sequence logos for the 5′TATA box (2.1a) and the 5′CCAAT box (2.1b) (well-conserved canonical eukaryotic transcription factor binding sites) are shown. The sequence logos were downloaded from JASPAR database (http://jaspar.genereg.net/) and represent consensus binding sites from 502 unrelated eukaryotic RNA pol II promoter regions. (Bucher et al, 1990). The letters on the X-axis represent the bases in order from the 5′ end. The Y-axis represents information content in bit score from 0 to 2.



2.1 a



2.1 b

Figure 2.1

Sequence degeneracy in known binding sites

Figure 2.2: Modular nature of promoters: Transcription factor binding sites are not positionaly aligned when compared across related promoter sequences. Different shapes indicate different binding sites for different proteins. The arrows indicate transcription start, "ATG" indicates translational start. Filled in shapes indicate that the binding site is present on the opposite strand.

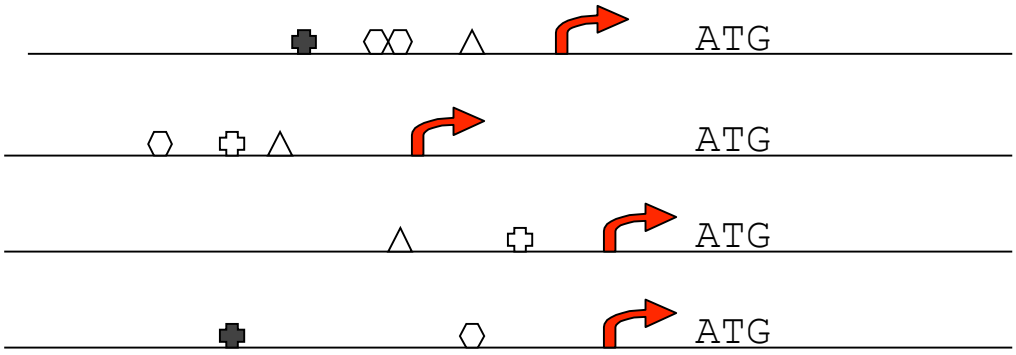


Figure 2.2

Modular nature of promoters

CHAPTER 3

IDENTIFICATION OF PUTATIVE *CIS*-REGULATORY ELEMENTS IN *CRYPTOSPORIDIUM PARVUM* BY *DE NOVO* PATTERN FINDING [1]

[1] Mullapudi N, Lancto C, Abrahamsen M and Kissinger J.C. 2007. *BMC Genomics.* 8:13-24. Reprinted here with permission of publisher.

Abstract

Background: *Cryptosporidium parvum* is a unicellular eukaryote in the phylum Apicomplexa. It is an obligate intracellular parasite that causes diarrhea and is a significant AIDS-related pathogen. *Cryptosporidium parvum* is not amenable to long-term laboratory cultivation or classical molecular genetic analysis. The parasite exhibits a complex life cycle, a broad host range, and fundamental mechanisms of gene regulation remain unknown. We have used data from the recently sequenced genome of this organism to uncover clues about gene regulation in *C. parvum*. We have applied two pattern finding algorithms MEME and AlignACE to identify conserved, over-represented motifs in the 5′ upstream regions of genes in *C. parvum*. To support our findings, we have established comparative real-time -PCR expression profiles for the groups of genes examined computationally.

Results: We find that groups of genes that share a function or belong to a common pathway share upstream motifs. Different motifs are conserved upstream of different groups of genes. Comparative real-time PCR studies show co-expression of genes within each group (in sub-sets) during the life cycle of the parasite, suggesting co-regulation of these genes may be driven by the use of conserved upstream motifs.

Conclusion: This is one of the first attempts to characterize *cis*-regulatory elements in the absence of any previously characterized elements and with very limited expression data (seven genes only). Using *de novo* pattern finding algorithms, we have identified specific DNA motifs that are conserved upstream of genes belonging to the same metabolic pathway or gene family. We have demonstrated the co-expression of these genes (often in subsets) using comparative real-time-PCR experiments thus establishing evidence for these conserved motifs as putative *cis*-regulatory elements. Given the lack of prior information concerning expression patterns and

organization of promoters in *C. parvum* we present one of the first investigations of gene regulation in this important human pathogen.

Background

Cryptosporidium parvum is an apicomplexan parasite that causes diarrhea in humans and livestock and is recognized as a common opportunistic and potentially life-threatening pathogen in AIDS patients. It is therefore considered a major public health problem (Spano and Crisanti 2000). *Cryptosporidium parvum* has a complex, obligate intracellular life cycle that is characterized by a series of asexual and sexual developmental stages. Infection is initiated by the ingestion of environmentally resistant oocysts that release sporozoites capable of invading intestinal epithelial cells. The obligate intracellular nature and complex life cycle make it difficult to study the developmental biology of the organism. Purification of the parasites from host cells is currently impossible. The different life cycle stages cannot be reproduced under *in-vitro* conditions (Abrahamsen and Schroeder 1999; Girouard et al. 2006). The situation is complicated further by the fact that the parasite is not amenable to either long-term cultivation or genetic dissection. Clearly, alternative approaches to investigate fundamental gene regulatory mechanisms in this important pathogen are required. Analysis of genomic sequence data and RT-PCR are two of the few available options. Genomes of two *Cryptosporidium* species (*C. parvum* and *C. hominis*) have recently been sequenced (Abrahamsen et al. 2004; Xu et al. 2004). The animal pathogen *C. parvum* has a highly A+T rich (70%), compact genome of 9.1 Mb comprising 8 chromosomes that are believed to encode 3952 protein coding genes separated by very short intergenic spaces of around 0.5 kb. Only 5-20% of the genes are thought to contain introns. Sequence analysis has revealed a reduced transcriptional and regulatory apparatus in comparison to other eukaryotes (Abrahamsen et al. 2004; Xu et al. 2004).

The study of promoters and *cis*-regulatory elements in apicomplexan parasites presents an interesting challenge. A few gene-specific experiments in the apicomplexan parasite

Toxoplasma gondii have revealed the absence of canonical elements such a TATA box in promoter regions. Instead, independent gene-specific studies have identified other motifs to be significant in gene expression (Mercier et al. 1996; Kibe et al. 2005; NF et al. 2006). In *C. parvum,* experimental analysis of promoters and gene expression, including microarrays, is currently not possible due to the aforementioned experimental limitations. However, the availability of two complete genome sequences and several bioinformatics tools to mine sequence data offer alternative approaches to identifying putative cis-acting promoter elements. We undertook a computational approach to identify conserved, over-represented DNA motifs in the intergenic regions of the genome that could serve as putative *cis*-regulatory elements. We have made an attempt to characterize regulatory elements in the absence of any known elements and limited expression data for a select number of genes. Following data mining analyses, we correlated our computational findings with independent experimental analyses (Figure 3.1). Our strategy involved grouping genes based on function and mining for conserved motifs in the upstream intergenic regions. We applied two pattern finding algorithms MEME (Bailey and Elkan 1994) and AlignACE (Hughes et al. 2000) to identify conserved, over-represented DNA motifs. We then employed comparative real-time PCR to establish the expression profiles of the genes examined.

Results

Genes with conserved upstream motifs have similar expression profiles

The groups of genes selected in this study are involved in parasite-specific functions as well as housekeeping-type activities (Table 3.1). Parasite-specific gene families included genes encoding cryptosporidial oocyst wall proteins (COWPs) and large secretory proteins (Cp LSPs), genes known to show concerted post-infection expression patterns (Abrahamsen et al. 2004; Templeton

et al. 2004). The housekeeping genes used in the analysis included genes involved in nucleotide salvage, DNA replication and glycolysis. We show that each of these groups of genes share different conserved upstream motifs. No common, general motif was conserved across all groups, barring AT-rich stretches, which were not statistically significant given the AT-richness of the genome.

Algorithms for pattern finding can report several hundred motifs ordered by their statistical significance as compared to a background model. Since statistical significance alone is not a sufficient indicator of biological significance, we applied a rule-based approach to identify candidate motifs that warrant further investigation. Candidate motifs were required to be within the top ten motifs predicted by both algorithms; display high information content and preferably show multiple occurrences within each upstream region. Information content as determined by MEME depends on the frequencies of the bases in a given column compared to the overall frequencies of those bases in the group of sequences. The more conserved the position is and the more rare the conserved nucleotides are, the higher the information content is. We provided MEME with a background file containing all inter-coding regions in the genome against which information content was calculated for each motif. (See Table 3.2 and Table 3.3 for the scores, E-values and information content of reported motifs).

Each group of genes that shared conserved, upstream motifs was examined for correlated expression profiles via comparative real-time PCR at 6 different post-infection time points. The housekeeping groups of genes were further resolved into three sub-sets based on their expression profiles. For each sub-set, a second iteration of pattern finding was performed to determine if conserved motifs within each sub-set existed. We find a correlation between genes that contain

distinct, conserved, upstream motifs and their corresponding expression profiles over a 72h post-infection period.

Genes encoding the COWP family

The cryptosporidial oocyst wall proteins comprise a multi-gene family that demonstrate a defined pattern of expression during *in vitro* development, with expression levels peaking at 48h through 72h post-infection (Templeton et al. 2004). Genes encoding members of this family are scattered throughout the genome and not clustered in a tandem array. The most significant motif found in the upstream regions of these genes as determined by both algorithms was a 12 bp motif (Figure 3.2a). This motif is present one or two times in the upstream region of all COWP genes, and when present in pairs, the motifs are often within 50 – 100 bp of each other. The promoter regions of this gene family are not alignable outside of the conserved motifs identified, indicating that the conserved motif is not simply a function of recent gene & promoter duplication.

Genes encoding the Cp LSP family

The large secretory proteins comprise a gene family that shows genomic co-localization in a cluster on chromosome 7. They are also co-expressed during the life cycle (Abrahamsen et al. 2004). Figure 3.2b shows a single DNA motif found upstream of each of these genes, with a well-conserved sequence. This motif occurs 2-3 times in all of the upstream regions, and is often located within –350 bp from the translational start. As is the case with the COWP gene family, the promoter regions of this gene family are also not alignable outside of the conserved motifs identified, indicating that the conserved motif is not simply a function of recent gene & promoter duplication.

Genes involved in nucleotide metabolism

Cryptosporidium parvum possesses highly streamlined nucleotide metabolic pathways, relying on the host cell for the salvage of both purine and pyrimidine residues. These pathways also contain genes that have been transferred into the nuclear genome of *C. parvum* from bacteria and plants via intracellular or horizontal gene transfer. The essential functions of these genes and their distinct evolutionary origin make them important drug-targets in developing anti-cryptosporidial chemotherapy (Striepen et al. 2004). We examined ten genes involved in nucleotide salvage and modification to identify significant motifs common to their upstream regions. We could not find a significant motif reported by both algorithms to be present upstream of all of the genes. MEME alone reported an 8 bp AT-rich motif present at least once in all the sequences at varying positions from the translational start (Figure 3.3a). Based on their real-time PCR expression profiles, the genes were divided into 3 sub-sets (sub-set 1, 2 and 3). Sub-set 1 contained three enzymes involved in the transport and modification of purines (AT, IMPDH and GMPS), and also one pyrimidine-modifying enzyme (CTPS). These genes were characterized by high expression levels at 2h and 12h post-infection and the most significant motif specific to this sub-set was a 10 bp motif shown (Figure 3.3a, sub-set 1). This motif was often found multiple times in the upstream regions and almost always found on the reverse strand. Three remaining pyrimidine-modifying enzymes (RDPR, dCMPD and DHFR-TS) had expression levels that peaked at 48h post-infection and dropped subsequently. They comprise sub-set 2. These genes contained a 12 bp motif in their upstream regions, also seen at varying positions from the translational start (Figure 3.3a, sub-set 2). The three kinases (AK, TK and UK) involved in nucleotide salvage were grouped together in sub-set 3 based on their high expression levels at 48h and 72h post-infection. They were found to contain a conserved 14-bp AT-rich motif in their upstream regions (Figure 3.3a, sub-set 3).

Genes involved in DNA replication

Analysis of the *C. parvum* genome reveals that the organism possesses a reduced complement of genes involved in DNA replication (Abrahamsen et al. 2004). We chose to study genes involved in DNA replication expecting that they would be co-regulated in a time-dependent manner associated with the life cycle (Spellman et al. 1998; Bell and Dutta 2002)**.** The most significant motif identified by both MEME and AlignACE was a single G-rich motif present upstream of all of the genes occurring multiple times in some of the upstream regions (Figure 3.3b). These genes could be resolved into 3 sub-sets based on their comparative RT-PCR expression profiles. Three genes peaking at 2h post-infection were classified into sub-set 1. A 14 bp motif with a core conserved 5′-CGCCAA-3′ sequence was found occurring once upstream of these three genes (Figure 3.3b, sub-set 1). At 6h post-infection, a few genes coding for MCM-like proteins and the single-stranded binding protein RP-A were found to peak in expression levels. These were classified into sub-set 2. The most significant motif found specific to this sub-set was an 11 bp motif occurring one or two times in the upstream regions of this sub-set (Figure 3.3b, sub-set 2). Most of the MCM-like proteins peaked at 48h post-infection and were classified into sub-set 3. These genes were found to contain a 13 bp motif, with a relatively less-conserved sequence, present one or more times in their upstream regions (Figure 3.3b, sub-set 3).

Genes involved in glycolysis

Glycolysis is considered to be the main source of energy in the Coccidia, and especially so in *C. parvum* due to the lack of evidence of a mitochondrion and a functional respiratory chain (Entrala and Mascaro 1997; Abrahamsen et al. 2004; Keithly et al. 2005). Ten genes associated with glycolysis were considered for this study, and a single motif was found to be over-

represented in all of their upstream regions (Figure 3.3c). This 10 bp motif contains a core 5′-GGCG-3′ sequence and is present multiple times in some of the upstream regions. No outstanding pattern with respect to the orientation or position relative to the start of translation is apparent. Comparative real-time PCR experiments resolved the glycolytic genes into three sub-sets based on their expression profiles during development. The genes that peaked at 6h- 12h post-infection were included in sub-set 1. The most significant motif found upstream of these genes was a 14 bp motif occurring uniquely, or many times and almost always on the opposite strand (Figure 3.3c, sub-set 1). Sub-set 2 is comprised of 3 genes exhibiting weakly correlated expression profiles, two of which peak at 48h post-infection. An 11 bp C-rich motif was found to occur one time in each of their upstream regions (Figure 3.3c, sub-set 2). We are encouraged that a conserved motif was found for this group, however, we remain unconvinced about the validity of this group based on their expression profiles alone since there is a discrepancy in the profiles. Sub-set 3 consists of 3 genes with expression levels peaking at 72h post-infection. A 9 bp motif was found to be common to their upstream regions. This motif was very well conserved at the sequence level (TGC[A/G][T/G]G[C/G]GA) (Figure 3.3c, sub-set 3) and was found occurring once upstream each gene.

Genome-wide occurrences of candidate motifs

The candidate motifs reported in this study were selected based on rules as described earlier. In all the cases except one, these motifs are also found elsewhere in the genome, upstream of other genes. This is not surprising given that these are short sequences, with several degenerate positions. Expression profiles are not yet available for every gene to test if the other genes containing these motifs upstream also display similar expression profiles. (Table 3.2).

Comparative studies in *C. hominis* and other apicomplexans

For each of the genes considered in this study (50 total) we retrieved the corresponding upstream regions from *C. hominis*. The intergenic regions between the two species are 95% identical. As expected, we could identify the exact same motifs in all upstream regions of the corresponding orthologs in *C. hominis*, except in cases where sufficient upstream sequence was unavailable due to unfinished genome sequence (9 genes, indicated by an asterisk in Table 3.1). Comparative analyses of upstream regions in *Toxoplasma gondii*, a more distant apicomplexan does not reveal the presence of the same conserved motifs in the groups studied (data not shown). This is not surprising considering the evolutionary distance between these species, and indicates that other apicomplexans may not serve as appropriate model systems for exploring the role of these *C. parvum* motifs.

Discussion

Eukaryotic gene regulation is a complex process that is regulated at various levels: epigenetic control via chromatin modification and reorganization; transcriptional control via proteins (transcription factors) that recognize specific signals in the DNA sequence (Struhl 1999); post-transcriptional regulation at the mRNA level (Day and Tuite 1998) and translational and post-translational control (Kozak 2005). We chose to look for conserved *cis* elements that may be representative of transcriptional regulation in *C. parvum* as this mechanism is most tractable to a *de novo* computational approach (to identify candidate motifs in the absence of any prior knowledge about the nature or organization of regulatory regions in this system). Our study was restricted to the 5′ upstream regions of each gene in consideration, and did not consider 3′ regions where *cis*-regulatory signals can also, presumably, exist.

There is significant evidence for transcriptional regulation in apicomplexan parasites. Microarray analyses in the apicomplexan parasite *Plasmodium falciparum* reveal a tightly controlled cascade of gene expression as reflected by the production of specific transcripts during the various erythrocytic developmental stages (Bozdech et al. 2003; Le Roch et al. 2004). Serial analysis of gene expression in another apicomplexan *Toxoplasma gondii* shows that unique stage-specific mRNAs are expressed during the course of its life cycle in the intermediate host (Radke et al. 2005). Recently, *T. gondii* has also been shown to contain a rich repertoire of chromatin and histone modifying enzymes found to play a role in stage-specific gene expression (Sullivan and Hakimi 2006). However, in both *T. gondii* and *P. falciparum*, (barring a few exceptions) co-expressed genes are not clustered within a region on a chromosome indicating that additional non-structural control mechanisms are involved in their regulation.

Cryptosporidium parvum is characterized by a compact genome (3952 protein coding genes in 9.1 Mb) and small intergenic regions (566 bp on an average). Genes are monocistronic and fewer than 20% of the genes are thought to contain introns (Abrahamsen et al. 2004; Xu et al. 2004), implying that gene-regulatory signals would likely be located in gene-proximal regions (Stamatoyannopoulos 2004). Previous studies of gene expression in *C. parvum* have examined genes clustered in the genome and those that are not (Deng et al. 2002; Abrahamsen et al. 2004; Templeton et al. 2004). Genomic clustering and co-expression has been observed in the *C. parvum* Large Secretory Proteins. The Cp LSP gene family exists as a cluster of seven adjacent genes on chromosome 7. These genes are co-expressed during *in vitro* development as shown by real-time PCR experiments (Abrahamsen et al. 2004). The co-expression of these clustered genes can be a function of shared control elements duplicated during expansion of the gene family or the result of epigenetic regulation. We have provided evidence for the existence of a

conserved upstream element (Figure 3.2b) that could possibly behave as a *cis*-acting signal to drive co-expression. Other groups of co-expressed genes are distributed throughout the genome (Deng et al. 2002; Templeton et al. 2004) indicating gene-specific control of expression.

Apicomplexan parasites still present a challenge when discussing mechanisms of *cis*-regulatory transcriptional control. Experimentally dissected promoters in *T. gondii* have not revealed the presence of known canonical eukaryotic promoter elements such as the TATA box. Independent gene-specific studies have revealed the presence of non-canonical regulatory elements in upstream regions of some genes in *T. gondii* (Mercier et al. 1996; Kibe et al. 2005; NF et al. 2006) and genome-wide studies in *P. falciparum* have indicated the presence of putative regulatory sequences correlated with expression profiles (Coulson et al. 2004; Callebaut et al. 2005). Preliminary genome-wide analyses of encoded proteins in various apicomplexans has revealed a reduced transcriptional machinery (Abrahamsen et al. 2004; Meissner and Soldati 2005). However, more than half of the predicted proteins in *C. parvum* (and other apicomplexan genomes) are hypothetical proteins. We hypothesize that the regulatory machinery in these parasites exists, but is so divergent that it cannot be identified by conventional similarity-based methods. Indeed, more sensitive, sequence-based search methods in *P. falciparum* have recently revealed the presence of basal transcriptional factors that were previously believed to be absent (Coulson et al. 2004; Callebaut et al. 2005). Other sensitive profile-based searches have reported the presence of a specific transcription factor ApiAP2 in *Plasmodium*, *Cryptosporidium* and *Theileria* spp. (Balaji et al. 2005).

The motivation for our study was to use genomic sequence information to infer the existence of putative *cis*-regulatory motifs in *C. parvum.* Most published methods used to identify *cis*-regulatory elements build upon *a priori* knowledge of regulatory structures or

expression patterns. Pattern finding algorithms are then trained, based on what is known about the organization and structure of regulons to identify additional elements. Such studies are not currently possible in *C. parvum* with traditional approaches like microarrays. We conducted a *de novo* search for conserved, over-represented short sequences in the upstream regions of genes that were grouped by metabolic function. We used two different algorithms and selected motifs predicted by both as extra evidence of significance. Both algorithms were provided with a background training set of all 3396 upstream intergenic regions in the genome to find statistically significant, over-represented motifs within the specified data sets (see methods). To determine if our findings had any functional significance, expression patterns for these genes were determined by real-time PCR experiments and the findings were correlated.

Our studies identified conserved upstream motifs that could possibly serve as recognition sites for hypothetical regulatory proteins in *C. parvum.* Biological sequences are non-random. The presence of conserved motifs in the upstream regions of genes that also demonstrate a similar expression profile indicates a possible biological function for these motifs. The actual biological role of these identified motifs remains to be determined. A possible function in splicing, post-transcriptional regulation and/or mRNA stability cannot be ruled out.

Unfortunately, given the current limits of the system, experiments focused on characterizing these functions cannot be performed. However, these motifs represent an exciting starting point to investigate the presence of specific trans-acting factors in *C. parvum* that may bind these *cis*-elements. We find that specific motifs emerge from different groups of genes studied, and no common motif across all the groups could be identified. It would be hard to believe that generalized transcription factors do not exist. One limitation of our method is the lack of transcriptional start site information for genes in *C. parvum* owing to the severe paucity

of EST sequences available. Aligning sequences based on transcriptional start site would be more informative with respect to revealing the presence of a possible global pattern present at a fixed location from the transcriptional start. Our study is also hindered by the AT-richness of the *C. parvum* genome (>70% in the intergenic regions). This biases the statistical significance of non A-T rich motifs as found over the background model. As more expression profiles are determined, the search can be enhanced by grouping genes based on their expression profiles into larger sets and searching for conserved patterns within.

We used two different pattern-finding tools to add to the evidence for selection of candidate motifs. MEME and AlignACE operate on different underlying algorithms and hence perform differently. In this study, the two programs displayed a fair degree of agreement in motifs reported in the top 3- 5 results. However, the best motif for each program rarely corresponded to the best motif as reported by the other program, and the motifs reported from both programs were rarely 100% identical. This is because of the inherent differences between the two algorithms and the criteria they employ to identify significant motifs. We used positional information to deduce results that overlapped between the two programs and picked candidate motifs accordingly.

The motifs identified in *C. parvum* could not be found in corresponding orthologs in more distant apicomplexan species such as *Toxoplasma*, indicating that other apicomplexans may not serve as a suitable model system for *C. parvum* in this regard. Indeed, the most pressing need is to develop better experimental techniques to test bioinformatics predictions in *C. parvum* itself. Our laboratory has applied this same method in the related apicomplexan parasite *T. gondii* where well-developed molecular genetic methods exist to transform parasites and carry out reporter expression assays. These expression studies have revealed a definite function for

some candidate motifs identified in this organism [Unpublished data, Mullapudi *et al*]. The study outlined here in *C. parvum* can contribute to the development of a database of "putative *cis*-regulatory elements" that will provide researchers with a starting point to investigate gene regulation in this parasite when the experimental tools become available. This resource would help alleviate the need for traditional "promoter-bashing" approaches and speed the progress of experiments aimed at characterizing transcriptional regulatory elements.

Conclusion

This is one of the first attempts to characterize *cis*-regulatory elements in the absence of any previously characterized elements and limited expression data. Using *de novo* pattern finding, we have identified specific DNA motifs that are conserved upstream of genes belonging to the same metabolic pathway or gene family. We have demonstrated the co-expression of these genes (often in subsets) using comparative real-time-PCR experiments thus establishing evidence for these conserved motifs as putative *cis*-regulatory elements. Given the lack of prior information concerning expression patterns and organization of promoters in *C. parvum*, the motifs identified here mark a starting point for the investigation of gene regulation in this important human pathogen.

Methods

Gene prediction and retrieval of intergenic regions

We used GLIMMER (Salzberg et al. 1999) to predict genes on the *C. parvum* genome, wrote scripts in PERL to extract gene-coordinate information and created the intergenic regions database (3396 sequences). To exclude the possibility of including coding regions in this set, a BLASTX was performed against known annotated *C. parvum* proteins using the set of intergenic regions as the query. 1000 sequences that contained portions of 100% identity to fragments of *C.*

parvum protein sequences were trimmed to remove the protein coding regions. The upstream region for a gene refers to the entire intergenic region upstream of the translational start.

Organization of genes into functional groups

In the absence of expression information, we classified genes into putatively co-regulated groups based on their function. To identify genes belonging to each pathway/ group, we made use of existing annotation, and BLASTP searches using orthologues from the related apicomplexans *Plasmodium falciparum* and *Toxoplasma gondii*.

Identification of conserved motifs in the upstream regions

We applied two pattern finding algorithms MEME and AlignACE to identify *de novo* patterns in the upstream regions. We used a background model based on the entire set of intergenic regions (3396 sequences) in *C.parvum* to train these algorithms. To identify patterns, the length range was set between 8 to 20 bp, and three different modes of occurrence were specified. The top 10 non-overlapping results from each algorithm were examined and compared, and the best motifs predicted by both algorithms were selected. We used WebLogo (Crooks et al. 2004) to create sequence logos to represent the best motifs found in each search.

***C. parvum* culture and RNA isolation**

C. parvum infected cultures - Human ileocecal adenocarcinoma cells (HCT-8, ATCC CCL-244; American Type Culture Collection, Rockville, MD.) were plated on 10cm plates and cultured to approximately 70% confluency in RPMI 1640 medium supplemented with 10% fetal bovine serum (FBS), sodium pyruvate, and antibiotics/antimycotic solution (100U penicillin G/ml, 100µg streptomycin/ml and 0.25µg amphotericin B/ml). *C. parvum* oocysts (Iowa strain) harvested from calves were purchased commercially (Pleasant Hill farms), stored at 4˚C and used for *in vitro* infections prior to three months of age as previously described (Templeton et al.

2004). Briefly, oocysts were surface sterilized by treatment with a 1:3 dilution of Clorox bleach (1 ml/ 3 x 10^7oocysts) on ice for 7 minutes, washed repeatedly with Hank's buffered saline solution (HBSS), and stored in Cp media [RPMI 1640 media containing 10% fetal bovine serum, 15 mM Hepes, 50 mM glucose, 0.1 u bovine insulin/ml, 35 µg ascorbic acid/ml, 1.0 µg folic acid/ml, 4.0 µg 4-aminobenzoic acid/ml, 2.0 µg calcium pantothenate/ml, 100 U of penicillin G/ml, 100 µg of streptomycin/ml and 0.25 µg of amphotericin B/ml (pH 7.4)] at 4˚C overnight. HCT-8 cultures were switched to Cp media approximately 18 hours prior to infection. Oocysts were warmed to room temperature for 30 minutes, and added to HCT-8 monolayers at a 1:1 ratio. Cells were incubated in a humidified incubator at 37˚C in an atmosphere containing 5% CO_2. Following a 2h excystation period at 37˚C, cells were washed repeatedly with warm HBSS and incubated at 37˚C in fresh Cp media until harvested. Infection was estimated to be between 70%-90% depending on the batch and storage period of oocysts. Total RNA was harvested in TRIzol reagent (Invitrogen) at 2, 6, 12, 24, 48, and 72 hours post infection and purified by following manufacture's instructions. Mock-infected cultures, cultures treated identically with the exception of infection, were harvested at exact time points as *C. parvum* infected cultures. Three independent time-courses were plated, infected, and harvested for this study.

Comparative real-time PCR

To investigate gene expression during *C. parvum in vitro* development, gene-specific primers (see additional file "S1-primers.xls" for primer sequences) were designed and used in comparative real-time PCR analysis. First strand cDNA was made using manufacturer (Invitrogen) protocols. Briefly, 2µg of total RNA that had been previously DNased following manufacturer instructions (Turbo DNase, Ambion) was mixed with 0.5µg of random hexamer and RNase-free water (to bring up the volume to 12µl), heated at 70˚C for 10 min, and cooled on

ice. To this mixture was added a 7µl aliquot, consisting of 4µl of 5X first strand buffer, 2µl of 100 mM dithiothreitol, and 1µl of 10 mM dNTP mixture. The reaction was equilibrated at 42˚C for 2 min on an iCycler (BioRad), after which 1µl (200 U) of SuperScript II RT was applied. The reaction mixture was then incubated at 42˚C for 50 min, heated at 70˚C for 15 min, and held at 4˚C. Identical reactions were set up without the addition of reverse transcriptase to test for the presence of contaminating genomic DNA using primers to *C. parvum* rRNA and 50 cycles of PCR under standard conditions. cDNA made from RNA containing no detectable product in the above PCR reaction was used for comparative real-time PCR.

Comparative real-time PCR was performed using a Stratagene Mx3000*P* real-time instrument in a 96 well format. Due to the sensitivity of the machine the amount of ROX normalizing dye needed to be reduced in the reactions. Therefore, 20µl reactions were set-up using a modified master mix consisting of 10µl 2X SYBR master mix (1 part SYBR green master mix containing ROX dye to 5 parts SYBR green master mix without ROX dye), 2ηmol of each primer and water to 13µl. cDNA was diluted 1:150 and 7µl of template was added to each reaction. After an initial denaturation at 94˚C for 2 min, the reaction mixture underwent 42 cycles of denaturation at 94˚C for 30 sec, annealing at 58 or 59˚C for 20 sec, and extension at 68˚C for 30 sec. Fluorescence was read after the end of each annealing cycle. Following the end of amplification cycles, a melting curve was run. This cycle started with an initial denaturation step at 94˚C followed by annealing at 56˚C. Melting was performed by increasing the temperature in single degree increments until the temperature reached 94˚C. Fluorescence of SYBR green was read at each increase and the data was plotted onto a graph using the Mx3000*P* software. cDNA made from RNA harvested from three independent timecourses was run in duplicate reactions and the average Ct value of each duplicate was determined using the Stratagene Mx3000*P* software. As

the number of developing *C. parvum* life stages within infected cells changes over time, primers specific for *C. parvum* 18S rRNA were used to normalize the amount of cDNA product of the target genes to that of *C. parvum* rRNA in the same sample. Due to the fact that rRNA is much more abundant than any specific mRNAs, the cDNA was diluted an additional 40 times and reactions were set up as above using three replicates of each time-course. Average *C. parvum* rRNA Ct values for each time-course was determined as above. An additional single time point of each time-course was run using *C. parvum* rRNA primers on each plate containing target genes to test the consistency of runs from plate to plate.

To determine comparative gene expression of a target gene, the average Ct values of all time points (expressed in log scale) of a single biological replicate were linearized and the ratio of expression for each time point was determined by dividing each by the product of the time point with the lowest expression (highest Ct). The ratio of expression for each time point of *C. parvum* rRNA expression was determined exactly as above. Relative target gene expression was determined by normalizing the ratio of the target gene expression to the ratio of *C. parvum* rRNA expression. Values from three biological replicates were imported into Prism (GraphPad Software). The mean and standard errors for each time point was determined using Prism's statistical package and the resulting graph was normalized to 100% maximum expression.

List of abbreviations used

RT – Reverse Transcription; MEME – Multiple Em for Motif Elicitation; AlignACE – Aligns Nucleic Acid Conserved Elements

Authors' contributions

NM and JCK designed the analysis and NM conducted the computational analyses. CAL conducted the comparative RT-PCR experiments. NM drafted the initial manuscript and JCK provided comments and critical revisions to the manuscript. JCK and MSA coordinated the study. All authors have read and approved the final manuscript.

Acknowledgements

We thank Jinling Huang for assistance with gene predictions and Abhijeet A. Bakre for helpful discussions to improve the text of the manuscript. Haiming Wang and Mark Heiges assisted with the retrieval of *C. hominis* sequences for the comparative analyses. We thank the reviewers for comments that greatly increased the clarity and quality of this manuscript.

References

Abrahamsen M, Schroeder AA (1999) Analysis of intracellular *Cryptosporidium parvum* gene expression. Mol Biochem Parasitol.

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G et al. (2004) Complete Genome Sequence of the Apicomplexan, *Cryptosporidium parvum*. Science 304: 441-445.

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.

Balaji S, Babu MM, Iyer LM, Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res 33(13): 3994-4006.

Bell SP, Dutta A (2002) DNA replication in eukaryotic cells. Annu Rev Biochem 71: 333-374.

Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B et al. (2003) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. Genome Biol 4(2): R9.

Callebaut I, Prat K, Meurice E, Mornon JP, Tomavo S (2005) Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes. BMC Genomics 6: 100.

Coulson RM, Hall N, Ouzounis CA (2004) Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. Genome Res 14(8): 1548-1554.

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14(6): 1188-1190.

Day DA, Tuite MF (1998) Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. J Endocrinol 157(3): 361-371.

Deng M, Templeton TJ, London NR, Bauer C, Schroeder AA et al. (2002) *Cryptosporidium parvum* genes containing thrombospondin type 1 domains. Infect Immun 70(12): 6987-6995.

Entrala E, Mascaro C (1997) Glycolytic enzyme activities in *Cryptosporidium parvum* oocysts. FEMS Microbiol Lett 151(1): 51-57.

Girouard D, Gallant J, Akiyoshi DE, Nunnari J, Tzipori S (2006) Failure to propagate *Cryptosporidium* spp. in cell-free culture. J Parasitol 92(2): 399-400.

Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol 296(5): 1205-1214.

Keithly JS, Langreth SG, Buttle KF, Mannella CA (2005) Electron tomographic and ultrastructural analysis of the *Cryptosporidium parvum* relict mitochondrion, its associated membranes, and organelles. J Eukaryot Microbiol 52(2): 132-140.

Kibe MK, Coppin A, Dendouga N, Oria G, Meurice E et al. (2005) Transcriptional regulation of two stage-specifically expressed genes in the protozoan parasite *Toxoplasma gondii*. Nucleic Acids Res 33(5): 1722-1736.

Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361: 13-37.

Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A et al. (2004) Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. Genome Res 14(11): 2308-2318.

Meissner M, Soldati D (2005) The transcription machinery and the molecular toolbox to control gene expression in *Toxoplasma gondii* and other protozoan parasites. Microbes Infect 7(13): 1376-1384.

Mercier C, Lefebvre-Van Hende S, Garber GE, Lecordier L, Capron A et al. (1996) Common cis-acting elements critical for the expression of several genes of *Toxoplasma gondii*. Mol Microbiol 21(2): 421-428.

NF VANP, Welagen J, Vermeulen AN, Schaap D (2006) The complete set of *Toxoplasma gondii* ribosomal protein genes contains two conserved promoter elements. Parasitology 133(Pt 1): 19-31.

Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS et al. (2005) The transcriptome of *Toxoplasma gondii*. BMC Biol 3: 26.

Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H (1999) Interpolated Markov models for eukaryotic gene finding. Genomics 59(1): 24-31.

Spano F, Crisanti A (2000) *Cryptosporidium parvum*: the many secrets of a small genome. Int J Parasitol 30(4): 553-565.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9(12): 3273-3297.

Stamatoyannopoulos JA (2004) The genomics of gene expression. Genomics 84(3): 449-457.

Striepen B, Pruijssers AJP, Huang J, Li C, Gubbels MJ et al. (2004) Gene transfer in the evolution of parasite nucleotide biosynthesis. Proceedings of the National Academy of Sciences, USA 101(9): 3154-3159.

Struhl K (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell 98(1): 1-4.

Sullivan WJ, Jr., Hakimi MA (2006) Histone mediated gene activation in *Toxoplasma gondii*. Mol Biochem Parasitol.

Templeton TJ, Lancto CA, Vigdorovich V, Liu C, London NR et al. (2004) The *Cryptosporidium* oocyst wall protein is a member of a multigene family and has a homolog in *Toxoplasma*. Infect Immun 72(2): 980-987.

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM et al. (2004) The genome of *Cryptosporidium* hominis. Nature 431(7012): 1107-1112.

Figure legends

Figure 3.1. Flow-chart illustrating methodology. Pattern-finding was carried out in two iterations, first *de novo* and a second time using information obtained from comparative real-time PCR expression profiles.

Figure 3.2. Motifs identified upstream of oocyst wall and large secretory proteins. (a) Upstream regions of genes encoding cryptosporidial oocyst wall proteins, and the occurrences of the most significant upstream motif shared by all of these upstream regions . The positions of the motifs are drawn to scale. All positions are shown with respect to the translational start. Solid black symbols denote a motif located on the reverse strand. Sequence-logo displaying the information content for the over-represented motif. (b) Upstream regions of genes encoding cryptosporidial large secretory proteins, and the occurrences of the most significant upstream motif shared by all of these upstream regions. Sequence-logo displaying the information content for the over-represented motif. Expression profiles for both families of genes were published elsewhere (Abrahamsen *et al.* 2004; Templeton *et al.* 2004).

Figure 3.3. Results of motif and expression analyses. (a) Motifs and expression profiles associated with genes involved in nucleotide metabolism. Schematic representation of the upstream regions are shown for each gene. The location of 4 different candidate motifs are indicated by the use of four different shapes. The single motif found in each gene of the group is indicated by a circle. The locations of three additional candidate motifs, each associated with a sub-set of sequences are indicated by the remaining shapes drawn on the upstream regions and as indicated to the left of each sequence logo. Solid black shapes indicate motifs found on the

reverse strand. Comparative real-time PCR profiles for sub-sets of each group of genes organized by expression profile over a 72h period are shown as sub-sets 1-3 Each sub-set is associated with a single candidate motif. (b) Motifs and expression profiles associated with genes involved in DNA replication. (c) Motifs and expression profiles associated with genes involved in glycolyis.

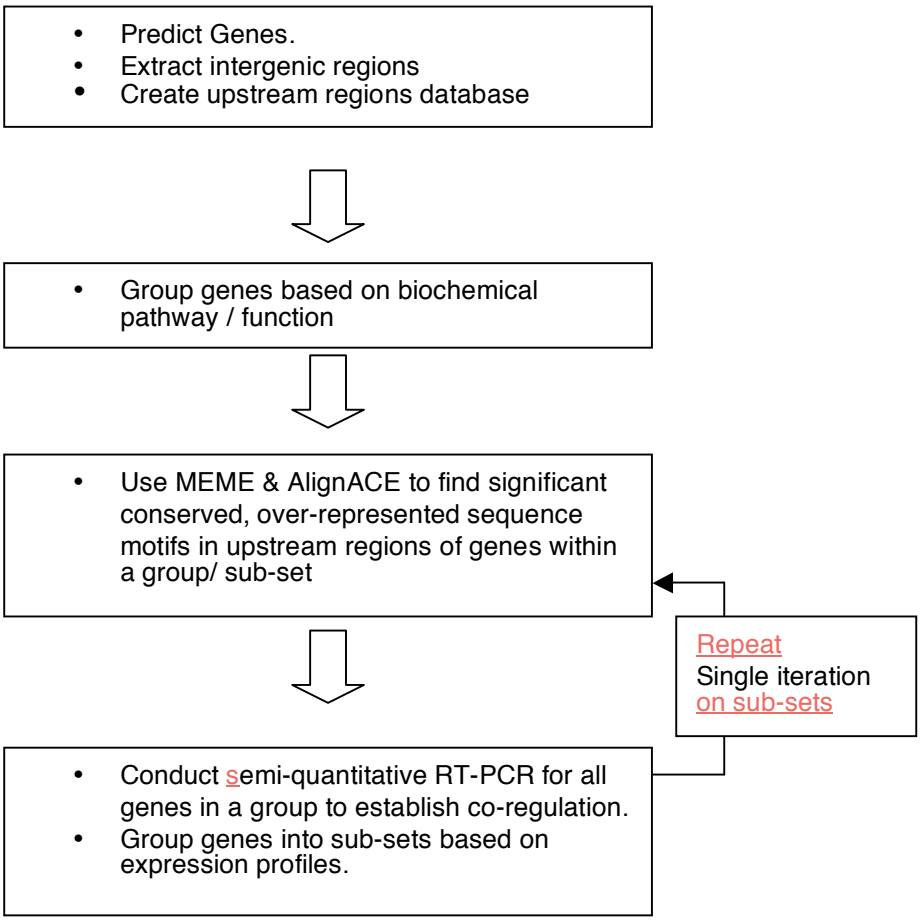


FIGURE 3.1

Flowchart illustrating methodology

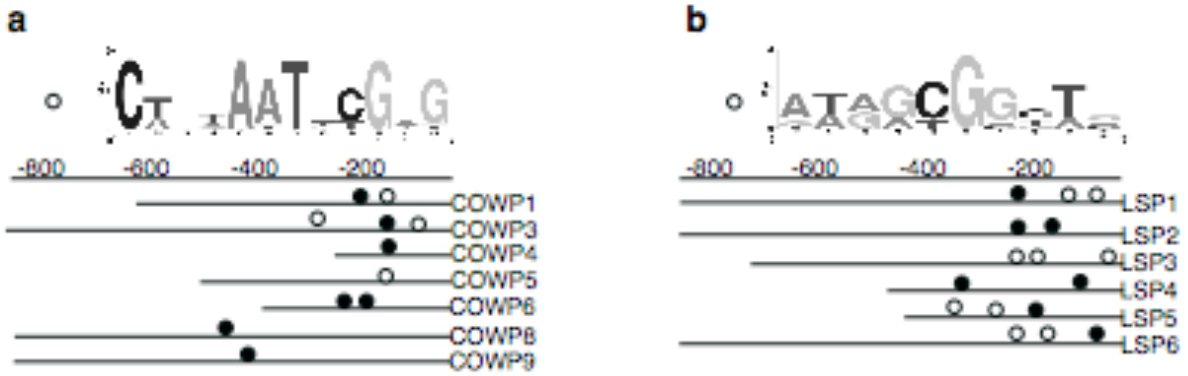


FIGURE 3.2

Motifs identified upstream of *C. parvum* oocyst wall and large secretory proteins

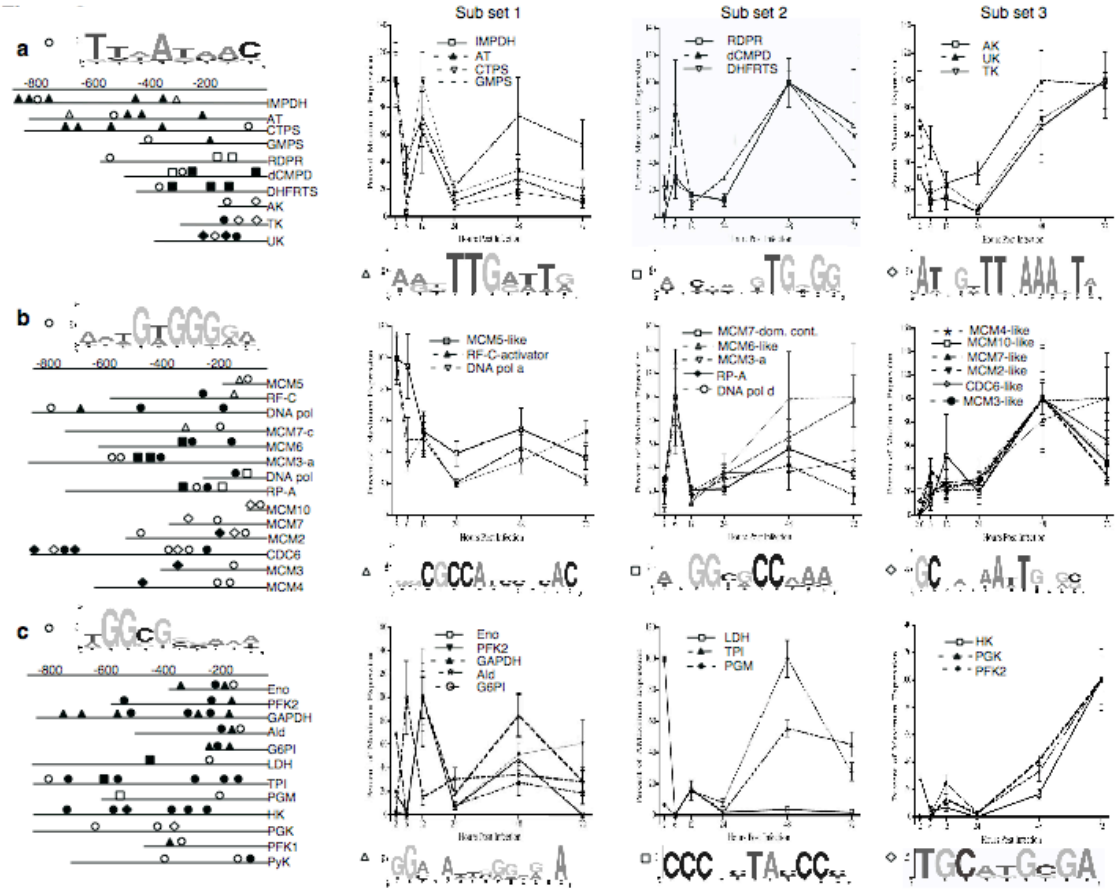


FIGURE 3.3

Results of motif and expression analyses for metabolic genes

Table 3.1 Genes used in this study

Descriptions of the genes used in this study, organized by functional group. The lengths of the respective upstream sequences (distance until the previous CDS) and annotated gene ids for each *C. parvum* gene are shown. Gene ids marked by a * are those for which corresponding ortholog information in *C. hominis* could not be obtained.

TABLE 3.1: List of genes used in the study

GENE NAME	ABBREV	LENGTH OF UPSTREAM	GENE ID
GLYCOLYSIS			
Hexokinase	HK	852	cgd6_3800
Phosphoglycerokinase	PGK	986	cgd7_910
Phosphofructokinase 1	PFK1	455	cgd3_1400
Phospphofructokinase 2	PFK2	575	cgd2_2130
Enolase	Eno	364	cgd5_1960
Glyceraldehyde-3-phosphate dehydrogenase	GAPDH	837	cgd6_3790
Fructose-bis-phosphate Aldolase	Ald	482	cgd1_3020
Glucose-6-phosphate isomerase	G6PI	225	cgd2_3200
Lactate Dehydrogenase	LDH	891	cgd7_480
Phosphoglucomutase	PGM	606	cgd7_4270*
Pyruvate kinase	PyK	718	cgd1_2040
Triose Phosphate Isomerase	TPI	500	cgd1_3040
NUCLEOTIDE SALVAGE			
Inosine monophosphate dehydrogenase	IMPDH	1065	cgd6_20*
Adenosine Transporter	AT	879	cgd2_310
Cytidine Triphosphate Synthase	CTPS	856	cgd5_1710
Guanidine Monophosphate Synthase	GMPS	423	cgd5_4520*
RibonucleosideDiphosphate Reductase	RDPR	627	cgd6_1950
deoxycytidine Monophosphate Deaminase	dCMPD	512	cgd2_2780
Dihydrofolate reductase-Thymidyl Synthase	DHFR-TS	481	cgd4_4460
Adenosine Kinase	AK	170	cgd8_2370*
Uridine Kinase	UK	404	cgd8_2810
Thymidine Kinase	TK	362	cgd5_4440
DNA REPLICATION			
DNA polymerase –□-subunit	DNA pol □	879	cgd8_870
DNA polymerase –□-catalytic subunit	DNA pol □	234	cgd6_4410*
MCM 10p-like	MCM 10p	92	cgd6_1710
MCM 2-like	MCM2	485	cgd2_1100
MCM3-associated	MCM3-a	882	cgd3_3570
MCM3-like	MCM3	370	cgd2_1600
MCM4-like	MCM4	638	cgd2_1250
MCM5-like	MCM5	155	cgd7_2920
MCM6-like	MCM6	604	cgd6_240*
MCM7-like	MCM7	374	cgd4_970
ORC/CDC6-like	CDC6	1301	cgd4_4320
RP-A ssb protein	RP-A	730	cgd2_4080
RAD24/RF-C activator	RAD24	571	cgd7_2660
MCM7 domain containing	MCM7-c	760	cgd8_3360
OOCYST WALL PROTEINS			

COWP1	COWP1	366	cgd6_2090
COWP3	COWP3	530	cgd4_670
COWP4	COWP4	197	cgd8_3350
COWP5	COWP5	170	cgd7_5150
COWP6	COWP6	688	cgd4_3090
COWP8	COWP8	604	cgd6_200
COWP9	COWP9	170	cgd6_210*
LARGE SECRETORY PROTEINS			
LSP1	LSP1	370	cgd7_3800*
LSP2	LSP2	370	cgd7_3810
LSP3	LSP3	225	cgd7_3820*
LSP4	LSP4	454	cgd7_3830
LSP5	LSP5	142	cgd7_3840
LSP6	LSP6	257	cgd7_3860
LSP7	LSP7	257	cgd7_3870

Table 3.2: Genome-wide occurrences of candidate motifs

This table gives the Information Content as reported by MEME for each candidate motif reported, the number of occurences of each candidate motif within it's respective sub-set. The motifs are also found elsewhere in the genome, as indicated in the last column showing genome-wide occurences of these motifs. Expression profiles are not yet available for every gene to test the deterministic nature of these candidate motifs in identifying other genes with similar expression profiles.

Table 3.2: Genome-wide occurrences of candidate motifs

GLOBAL				
GROUP	MOTIF	I.C (MEME)	SUBSET-SPECIFIC OCCURENCES	GENOME WIDE (IN 3396 UPSTREAMS)
			(# TIMES / # GENES)	(#TIMES /# GENES)
NUCLEOTIDE METABOLISM	TtnAtaac	7.1 bits	10/ 10	3397/3192
DNA REPLICATION	aatgTggggaa	14.3 bits	25/14	253/238
GLYCOLYSIS	GGcgggaaA	20.1 bits	28/ 12	1635/1261
COWP	ctctAatacgaga	15.5 bits	12/ 7	77/60
LSP	atagCGgctg	15.5 bits	18/ 7	1313/972
		SUB SET 1		
NUCLEOTIDE METABOLISM	agtTTGAtTg	18.9 bits	15/ 4	841/677
DNA REPLICATION	ggCGCCAtccncAC	26.9 bits	3/ 3	9/ 9
GLYCOLYSIS	tGaaAttgggaanA	23.1 bits	11/ 5	358 /334
		SUB SET 2		
NUCLEOTIDE METABOLISM	CacnatgTGcGgg	19.4 bits	8 / 3	642/568
DNA REPLICATION	AaggCgCcAaa	29.8 bits	7 /5	487/459
GLYCOLYSIS	CCCncTAcCCc	19.4 bits	3/ 3	3/ 3
		SUB SET 3		
NUCLEOTIDE METABOLISM	ATtgtTTnAAAAnTa	20.8 bits	6 /3	437/380
DNA REPLICATION	GCaaaaATTg	18.4 bits	10/ 6	392/363
GLYCOLYSIS	TGCatGcGA	16.2 bits	3/ 3	17 / 17

Table 3.3: Top-ten motifs identified in each search

Top scoring motifs found by both MEME and AlignACE. Candidate motifs reported in Figure 3.2 and 3.3 are denoted in bold. Note that the motifs found by the two programs are never identical. We used positional information to deduce overlapping motifs, and used the motifs identified by AlignACE to represent the consensus since AlignACE found more than one occurence of the same motif within the same sequence and hence produced a more degenerate motif.

Table 3.3 Top 10 motifs identified in each search

	MEME	E-VALUE	ALIGNACE	SCORE	MEME	E-VALUE	ALIGNACE	SCORE
	GLOBAL MOTIF				SUBSET 1			
NUC. METAB								
1	Taaaaccaac	4.80E+09	aaagaaatGg	15.1883	**agtTTGAtTg**	9.00E+04	**aatTTGattg**	11.2461
2	**TtnAtaac**	8.70E+10	AAaAAaAAaA	11.2017	GAgAtTTTg	6.60E+05	AAAaAaAAAA	8.49071
3	aAattcaAac	4.00E+11	aanggaGGaA	7.82835	ACAgAaaTnn	1.40E+06	AagcttgGTa	7.48283
4	tTTtaaaaAt	3.60E+12	aaangggGca	6.43035	AtTTaTTTct	2.50E+06	aTtTgAatga	5.70644
5	TtTaTAatta	4.60E+14			AAtgtTTngT	1.20E+07		
6	TtaattttTT	1.70E+15			TTTanAAgnn	1.70E+07		
7	atatatattt	3.40E+17			TAatAatgaA	2.30E+07		
8	aatattaaat	3.2E+18			AAntAagnag	2.60E+07		
9	tcanattcta	1.3E+21			cAacntnaTT	2.70E+07		
10	attttaanaa	9E+23			ATTTTATncn	4.90E+07		
DNA REP								
1	ttaaaaaaaa	1.70E-07	**aatGtGGgga**	51.5459	**ggCGCCAtcc**	1.00E+01	gAAAaaaAAa	12.7146
2	ttGgcgccaa	3.00E-02	ggcGCCaaaa	50.5833	cAaTTGCcC	2.30E+03	**ggCGCCAtcc**	12.0322
3	**aatgTgggga**	5.20E-01	AaaaAAAAaa	26.8671	cgGaAAAnAn	1.40E+04	gaaaAagaaa	5.66741
4	ggaaAaagaa	1.40E+16	gGcgccAaat	20.9919	CcCtcAgcT	1.40E+04	AaAnaAataa	1.66297
5	aaTTAaaA	7.30E+16	atggcGgaAa	9.79455	AagCtGcA	3.40E+05		
6	atttgnnTtA	6.10E+18			cATTTnGtGn	1.80E+05		
7	aatttAanaa	1.20E+20			TTggctgaAA	5.30E+05		
8	Aaataaaa	6.50E+21			AAgCAAnTtg	6.60E+05		
9	aaaatanatt	7.10E+22			CgCnTTTA	1.20E+05		
10	gaaaaanaat	1.50E+29			gnAGnanTag	1.90E+07		
GLYCOLYSIS								
1	**GGcgggaaA**	9.00E+05	**tGGcgggaaa**	47.7588	gGcaggaatA	5.70E+01	**gGaaattggg**	18.6463
2	tacaaatttc	1.00E+07	tGgcGcgaaa	24.6653	**tGaaAttggg**	2.30E+02	GGCGcnanaa	16.8582
3	AgAAAaAa	3.50E+11	aattTGcatG	19.6727	TgAGTganGA	2.40E+03	atAttagatA	13.047
4	tTttgaAAaa	1.7E+12	ttnanggaaa	17.2535	ggctaatggg	6.00E+04	nttGGCgnga	11.7294
5	taaaaaaAAA	2.20E+13	gngnnnaaan	16.9816	gagnctgnac	1.80E+05	tTtnttttaA	10.723
6	anttntaatn	9.7E+13	aataaAAAaA	13.4838	AtTTnGaagT	7.20E+05	ctggcggGcg	8.20166
7	TttAtTaAgt	4.3E+14	gcngGcgcna	13.2493	ATTTgaGA	3.40E+06	gcantantat	6.18004
8	AAAtAaTT	2.6E+15	ctaattggcg	12.1617	AaATangngc	1.70E+07	gAAntAgcaa	5.11496
9	TtaTATtt	6.8E+18	gttggcggga	10.8252	tAntaTnCcc	2.40E+07	AAAaaaATTt	4.41773
10	aaattngcgn	1.5E+20	anttatatgc	10.1912	AaaTAAATA	6.50E+07	ntgggCnggn	3.90505

Table 3.3 continued

	MEME	E-VALUE	ALIGNACE	SCORE	MEME	E-VALUE	ALIGNACE	SCORE
	SUBSET 2				SUBSET 3			
NUC. METAB								
1	**Cac?atgTGc**	24	**aactatgTGg**	14.5793	**ATtgtTTnAA**	1.90E+03	**AttgtTTnAA**	7.97953
2	cggtACC?Ca	5.50E+03	caantagtgc	14.01	ctCTTTTncC	2.00E+03	TtaaanTtta	4.19699
3	gTGctcGT	1.30E+04	Aaatggggca	9.89848	aTCcCaCT	3.70E+03	TnnaAgtGGG	1.43498
4	cTgagTG?GG	1.80E+04	tGtGGngant	8.88066	TTtCAATC	1.10E+05		
5	CCCCACcg	2.50E+04	ataagtGGcg	7.03635	gcagATACT	2.80E+05		
6	TGTggCAA	1.50E+05	Aanttgccac	5.26198	ccACTaTTc	4.30E+05		
7	Ga?ATTctCA	2.70E+05			GgtAtatTG	9.80E+05		
8	TcACattCTa	3.00E+05			gccgTAgaT	1.20E+06		
9	caTcTTTTtg	4.30E+05			cTaagAcAA	2.40E+06		
10	TTgTct?aTg	4.90E+05			AAttTtGAng	5.10E+06		
DNA REP								
1	**AaggCgCcAa**	4.60E+00	**aaGGcgCCaa**	25.2269	**GCaaaaATTg**	4.80E-05	**GCaaaaAtTg**	20.9322
2	AAagAaAAAt	7.70E+03	aaaaaGaGgg	24.5642	TttTtggcgC	2.40E-01	ggnGngaAnt	20.4301
3	gngaAaatGc	3.10E+04	AnaAAAaAaa	18.1664	taAtgcanct	7.00E+04	agtGtGGgga	19.9657
4	AAAgngAnAa	3.30E+04	Aaaataaaaa	16.903	Taaatannan	1.90E+09	tagGcGccaa	13.0036
5	CTTnCnTT	2.80E+05	aaAaaAaaAa	12.7821	AagtattaAn	1.9E+11	tGtgnGgaga	11.2129
6	TTTCcCnc	5.40E+05	GGcGCCaaan	10.9802	gnanatttta	1.5E+11	AatcAAAAAa	9.2768
7	ggaggAaTTt	5.40E+05	attaGGcgcc	6.12686	attAAnTaAt	1.5E+12	ttGgtgccaa	7.80461
8	GcAtAatTgt	2.10E+06			taaTtaAtan	1.8E+10	gAaantTtGn	7.51254
9	gTAaaAnaaA	2.60E+07			nttactaaAg	2.2E+13	aagtgggtaa	7.23736
10	TTATTcngTn	1.50E+06			aAtAnttaaa	1.5E+14	attgaaGcaa	6.16794
GLYCOLYSIS								
1	**CCCncTAcCC**	1.20E+02	gaaaanaana	17.1085	ngtttTCcCg	2.2	tgaggngcna	30.531
2	G?AaacTt?c	5.80E+02	**CCCncTAccC**	13.9919	**TGCatGcGA**	3.70E+01	**tTGcatgtga**	21.6646
3	GTggGcTc	1.30E+04	aaaaaAAaat	11.2284	TGaAagtggt	1.30E+03	ggcgagtaaa	18.2195
4	CcaG?TAgtC	6.80E+04	gggaaaaaAa	9.9098	AngTtgcatc	2.00E+03	GcagGngaaa	17.9725
5	GGa?AgtaAa	1.10E+05	aaaattntcg	9.47073	Gcanatntcn	2.20E+03	ggagggaaaa	17.2433
6	TTGCAAgAAt	1.20E+05	AtTtaaaaat	7.84612	gTAGGnaAAt	2.20E+04	aatagaGaaa	16.9023
7	gtGaGGATta	1.30E+05	AannAAAaAa	7.72311	AGtGCaaAnn	3.80E+04	gtgggGaaaa	15.961
8	cGgaAtat?A	3.10E+05	tGcaaGtaAa	4.15168	TGnAtnAntn	5.60E+05	taaattAGnA	13.3231
9	CaATattCCa	1.40E+06	Ggcgcntaan	2.40914	AnnangAAAt	1.70E+06	AgggaaaaAA	12.192
10	ACATt?cTAa	2.30E+06	TtaGattaca	1.93517	gnnttanATt	2.00E+06	AaaaAaAtnC	11.9477

Table 3.3 continued

COWP	MEME	E-VALUE	ALIGNACE	SCORE
1	**ctctAatacgaga**	1.30E+05	AataAtAaAAaAAanaA	14.4211
2	AAaaaaAgnaaatactgatt	7.40E+06	**CtctAaTtcGag**	11.6516
3	ataagtaanaAAaTactnattannaaat	2.20E+07	cttnaaaattaGaAttC	4.95359
4	tnacatnAAaAaTtg	3.30E+08	TtctcGCnaaA	4.52948
5	aTnanantttnnatgnattncntg	4.10E+09	ttctaGnaAAAatta	4.43072
6	taTTAtTg	5.60E+10	cattAaaAnTtG	4.39155
7	ttTangtttA	1.80E+12		
8	ctngaaaaaaTtatant	4.30E+12		
9	attnAacananaaat	7.30E+15		
10	tgtAattcanttatnnaaanT	3.50E+17		

LSP	MEME	E-VALUE	ALIGNACE	SCORE
1	**atagCGgctg**	2.10E-06	**atagcGgctg**	28.9734
2	aatAaTtgatgtAatctAtaAgtagaanaang	5.60E-03	ctGctatatgtTcaaaA	28.2041
3	AaTTTTttaAtagCn	3.80E+02	attgcggcTtaattatt	24.7394
4	gAAgaaAATttaATA	6.40E+04	AtaaaganaatAAtgtg	22.4015
5	ActGntatatGaTcaAaanncTttnnatc	2.60E+00	TtaaataataTaaaaar	16.872
6	AAtcTtAcA	1.60E+09	gCAtGcatAntaaA	16.2253
7	TTatATaaTAA	1.70E+08	aatGcaatnnnAtcAa	15.6644
8	AaATtcAT	3.60E+09	AaagtGgtaatacaatA	15.1515
9	aTAaATcT	1.20E+09	ttaaATCtAAa	13.8035
10	tgcAgtTt	4.90E+08	AtAgtaTtnaantacta	12.5213

CHAPTER 4

IDENTIFICATION AND FUNCTIONAL CHARACTERIZATION OF *CIS*-REGULATORY ELEMENTS IN THE APICOMPLEXAN PARASITE *TOXOPLASMA GONDII*[1]

[1] Mullapudi N. and Kissinger J.C To be submitted to *Genome Biology*.

ABSTRACT: *Toxoplasma gondii* is a member of the phylum Apicomplexa; a phylum that consists entirely of parasitic organisms that cause several diseases of veterinary and human importance. The fundamental mechanisms of gene regulation in this group of protistan parasites remain largely uncharacterized. Owing to their medical importance genome sequences are available for several apicomplexan parasites. Their genome sequences reveal an apparent paucity of known transcription factors and the absence of canonical *cis*-regulatory elements. We have approached the question of gene regulation from a sequence perspective by mining the genomic sequence data to identify DNA motifs conserved upstream of genes that can serve as putative *cis*-regulatory elements. We have subsequently characterized the function of some of these conserved elements in reporter assays in the parasite, and show a sequence-specific role in gene-expression for 5 out of 8 identified elements. This work demonstrates the power of pure sequence analysis in the absence of expression data or *a priori* knowledge of regulatory elements, and its applicability towards understanding transcriptional regulation of systems where little is known.

INTRODUCTION:

Toxoplasma gondii is an obligate intracellular parasite belonging to the phylum Apicomplexa. The parasite exhibits a complex developmental life-cycle wherein it is capable of switching between a rapidly dividing tachyzoite form and a quiescent bradyzoite form within the asexual stage of its life-cycle (Black and Boothroyd 2000). During its asexual stage, it exhibits a wide host range, capable of infecting a variety of warm-blooded animals. Infection is of greater concern in AIDS or immuno-supressed patients, where it can lead to neurological, mental and ocular defects. It is also responsible for human birth defects or spontaneous abortions caused by trans-placental transmission in infected pregnant women (Alexander et al. 1997; Gilbert et al. 2001). Given it's wide host-range and medical importance, understanding fundamental processes of gene regulation is important for developing methods aimed at controlling infection and disease.

Regulation of gene-expression is an important requirement for parasitic lifestyle. There are many levels at which organisms can control gene-expression, and these include chromatin-mediated epigenetic regulation, transcriptional regulation, and post-transcriptional and post-translational regulation (Mattsson et al. 1997; Struhl 1999). Very little is known about how *T. gondii* and other apicomplexan parasites regulate their genes. A relatively small number of gene-specific studies in *T. gondii* have identified non-canonical *cis*-regulatory elements that were found to play a role in downstream gene expression (Soldati and Boothroyd 1995; Mercier et al. 1996). Preliminary surveys of the complete genome sequence have revealed a paucity of known specialized transcriptional factors encoded in the genome (Meissner and Soldati 2005). Recent studies have focused on dissecting the developmental signals responsible for inter-conversion between the tachyzoite and bradyzoite developmental stages and preferential gene expression

that characterizes these stages. To this end, the study of stage-specific genes and their promoters (Ma et al. 2004; Kibe et al. 2005) has revealed the presence of regulatory elements in the promoter region that are responsible for preferential gene expression in different life-cycle stages. Large-scale analyses of gene expression from key developmental life-cycle stages (Radke et al. 2005) point to the absence of chromosomal clustering of co-expressed genes, and the presence of unique stage-specific mRNAs in each developmental stage. However, promoter organization and the presence of specialized transcription factors for the recognition remain largely unexplored areas.

The medical importance and evolutionary divergence of the apicomplexan parasites has motivated a rapidly growing collection of genome sequencing efforts for this group. Sequence information is an excellent starting point to identify *cis*-acting signals in the genome and to uncover the underlying gene regulatory mechanisms. Sequence analysis to identify conserved *cis*-regulatory signals is typically augmented by at least one of two types of information: the organization of regulons and known sequences of conserved transcription factor binding sites, or large-scale gene expression information (e.g. from microarray studies), that provide data sets of co-regulated genes within which conserved transcription factor binding sites can be identified (Hughes et al. 2000). Known canonical eukaryotic *cis* elements have not yet been reported in *T. gondii*. Instead, other novel elements have been found to play a role in gene expression (Mercier et al. 1996). In the absence of this starting information, we have adopted a *de novo* approach to identify conserved sequence elements that could serve as putative *cis*-regulatory elements. We have then experimentally verified the role for these candidate elements in the parasite, establishing their role in gene expression. Our study includes four different groups of genes that share parasite specific or metabolic functions in the parasite. We describe a computational

framework for the identification of novel *cis*-regulatory elements in eukaryotic non-model systems.

MATERIALS AND METHODS:

Computational analyses:

Whole genome sequence and gene-predictions for *Toxoplasma gondii* were obtained from toxodb.org (Kissinger et al. 2003). Scripts were written in PERL to extract the upstream sequence (2 kb or until the previous gene, whichever was smaller) for every predicted gene to create an "upstream sequence database". This database was screened for possible protein coding regions to eliminate mis-annotations by the gene predictions by performing a BLASTX against all protein sequences in the non-redundant database at NCBI, and sequences that contained significant similarity to coding sequences were pruned. Existing annotation and BLAST-derived sequence similarities were used to assign gene function and classify genes into groups based on function. The pattern-finding algorithm MEME (Bailey and Elkan 1994) was used to identify over-represented conserved motifs in the upstream regions of genes within a group. MEME was run using the parameters minw = 8, maxw = 20, in all three modes (tcm, oops and zoops) and the results were manually examined to pick candidate motifs. Upstream sequences of corresponding orthologs from *Eimeria tenella* (the closest coccidian parasite genome sequence available) were also used whenever possible to identify evolutionarily conserved motifs. In order to narrow down the results of MEME, a rule-based approach was adopted to pick candidate motifs for subsequent experimental validation. Candidate motifs included those that were over-represented in the upstream regions in comparison to the background set (entire upstream regions database); showed considerable conservation in sequence and/or position and were present in all the sequences within each group.

Molecular techniques:

For each group of functionally related genes considered in this study, a promoter that contained a single occurrence of the candidate motif was chosen to study the role of the motif in driving expression. Promoter sequences were PCR-amplified from parasite genomic DNA and a two-step overlap-extension PCR technique was employed to carry out site-directed mutagenesis (Sambrook et al. 1989) to change the candidate motif sequence. All or majority of the bases in each motif were substituted via transversions or by a polyA tract thus destroying the original sequence of the candidate motif but maintaining the spacing within the promoter (Figure 4.1-4.4). Wherever possible, a restriction enzyme site was introduced to facilitate convenient screening of the mutagenized PCR product. Successful mutagenesis was confirmed by sequencing or by restriction digest analysis. The Gateway™ cloning system was used to clone the wildtype and mutagenized promoters individually upstream of a firefly-luciferase expressing vector. As an internal control, a constitutive promoter (*T. gondii* α-tubulin promoter)-driven renilla-luciferase expressing construct was co-transfected along with the experimental construct.

Parasite culture and transient transfections:

Toxoplasma gondii RH tachyzoites were cultured in human foreskin fibroblast (hTERT cells, BJ Biomedicals) as previously described (Soldati and Boothroyd 1995). Transient transfection was performed via electroporation, using freshly lysed parasites, needle-passed and filtered through a 3 μm filter and resuspended in cytomix (Soldati and Boothroyd 1993, 1995). Immediately prior to use, freshly prepared 2 mM ATP and 5mM glutathione were added to the cytomix and sterile-filtered. For each co-transfection, 2 X 10^7 parasites were transfected via electroporation with a mixture of sterile circular plasmid DNA of α-tub-renilla (control) and test-firefly mixed in a ratio of 2.5 : 1.0 (40 μg test + 16 μg control). Electroporation was performed

in a 2mm gap cuvette using a BTX electroporator: 1.8 kv, 100O, 25uF). Post-electroporation, the parasites were allowed to rest for 15 minutes in the cuvette and then transferred to T25 tissue culture flasks. 18-24 h post-electroporation, the cells were scraped and lysed using passive lysis buffer (Promega) and a dual luciferase assay was performed using the Promega Dual Luciferase Kit (DLR). Briefly, the different substrate requirements for each enzyme, renilla luciferase and firefly luciferase allowed us to assay enzyme activity sequentially within the same extract. Enzyme activity was measured using a DLR-ready luminometer. Each electroporation experiment was performed in triplicate and luciferase assays were performed in duplicate for expression measurements. The Students T-test was used to calculate significant differences between wild-type and mutagenized promoter activity; $p < 0.005$ was considered significant.

RESULTS AND DISCUSSION:

We analyzed four different functional groups of genes for the presence of conserved, over-represented upstream motifs within each group. We find that different groups of genes share different over-represented motifs, no global motif emerges from our studies to be shared by all groups. We present here the results of pattern-finding and accompanying experimental evidence to establish the biological role of the motifs considered in this study.

Genes involved in glycolysis:

Toxoplasma gondii, like *E. tenella* and *C. parvum* uses glucose as its main source of energy in its rapidly dividing tachyzoite stage (Denton et al. 1996). Phylogenetic analyses have shown that two of the glycolytic genes in *T. gondii*: enolase and glucose-6-phosphate isomerase are closely related to their corresponding homologs in plants, suggesting that they were acquired and making them potential drug targets due to their distinct evolutionary origin (Dzierszinski et al. 1999). Glycolysis has also been actively studied with respect to stage-differentiation in *T.*

gondii. Three key glycolytic enzymes: glucose-6-phosphate isomerase, lactate dehydrogenase and enolase exhibit developmentally regulated expression (Dzierszinski et al. 2001). Stage-specific cDNAs have been isolated that encode for distinct isoforms of lactate dehydrogenase: *LDH1* (tachyzoite) and *LDH2* (bradyzoite) (Yang and Parmley 1995). Experimental evidence based on the detection of their respective mRNA and protein products indicates that *LDH1* is post-translationally repressed while *LDH2* is transcriptionally induced in bradyzoites (Yang and Parmley, 1997). Similarly, stage-specific cDNAs have also been isolated for distinct forms of enolase: *ENO1* (bradyzoite) and *ENO2* (tachyzoite) (Manger et al. 1998). It has recently been demonstrated that stage-specific expression of the two enolases is brought about by the presence of specific *cis*-regulatory elements in the promoter regions of these genes (Kibe et al. 2005). The regulation of the genes involved in glycolysis presents an intriguing case study from developmental, evolutionary an regulatory perspectives.

We analyzed the upstream sequences of eleven genes involved in tachyzoite glycolysis to identify conserved, over-represented sequence motifs. We report the analysis of two candidate motifs here: motif A, also found upstream of some orthologs in *Eimeria tenella* and motif B, found exclusively in *T. gondii*. Motif A, represented by the consensus 5′GCTKCMTY (Figure 4.1a) is an 8 bp well conserved sequence occurring at least once per sequence on the forward strand and sometimes also on the reverse strand. It does not show significant positional conservation, but motifs found upstream of orthologs in *E. tenella* are found to be 100% conserved in sequence to their counterpart in *T. gondii*. Motif A is exclusive to the upstream regions of the tachyzoite isoforms of the stage-specific glycolytic genes. Motif B is also an 8 bp motif represented by the consensus sequence 5′TGCASTNT (Figure 4.1a), with 6 of 8 bases conserved in more than 90% of the occurrences. This motif is present only once per sequence,

and is independent of orientation (Figure 4.1a). Motif B was also found in the upstream regions of the bradyzoite specific copies of enolase and LDH (Data not shown).

Mutagenesis of motif A to the sequence 5′AACAAACA in the *eno2* promoter resulted in a small but significant decrease ($p < 0.005$) in promoter activity. Mutagenesis of motif B to the sequence 5′CAACACAC within the *eno2* promoter resulted in no significant change in promoter activity (Figure 4.1b). We postulate that motif A plays a positive role in the expression of glycolytic genes in the tachyzoite stage. The strong evolutionary conservation of this motif in *Eimeria tenella* lends support to its functional significance. Motif B did not have a statistically-significant effect on promoter activity when mutagenized. The *Eno2* promoter appears to be an inherently weak promoter, so it is possible that small effects are not noticeable by this assay. It is also possible that this motif does not affect promoter activity in a sequence-specific way. The role of this motif in the bradyzoite-specific genes warrants further investigation.

Genes involved in nucleotide biosynthesis and salvage:

Purines and pyrimidines are the building blocks of nucleic acids in all living cells. All protozoan parasites are unable to synthesize purines *de novo* and depend upon salvage enzymes to obtain these from the host (Berens et al. 1995). Most protists, however, possess a full set of *de novo* pyrimidine biosynthesis enzymes, one of the exceptions in this case being *C. parvum*, which has evolved to also salvage pyrimidines from the host cell (Striepen et al. 2004). Enzymes involved in nucleotide metabolism in protozoan parasites can serve as promising drug targets because they are essential to the parasite's survival and are also evolutionarily distinct in some cases (Striepen et al. 2004). In *T. gondii*, it was found that *de novo* pyrimidine biosynthesis is essential for the virulence of the parasite (Fox and Bzik 2002). We examined eight genes encoding for enzymes involved in nucleotide biosynthesis and salvage in *T. gondii* and selected

two conserved motifs found in their upstream regions as candidates for experimental validation. Motif A is an A-rich 9 bp motif represented by the consensus 5′GCAAAMGRA (Figure 4.2a). Motif A was found to be very well conserved in 4 orthologs in *E. tenella*. This motif is present only once upstream of each gene and is always found on the positive strand. It is primarily located at 1000 to 1500 bp upstream of the translation start (Figure 4.2a). Motif B is an 8 bp long T-rich motif and is exclusive to *T. gondii*. It is represented by the consensus sequence 5′TTTYTCGC (Figure 4.2a) and is also found only once upstream each gene on the forward strand. The two motifs are typically present with 300 to 400 bp of each other. The occurrences of motifs A and B upstream of the eight genes are shown in Figure 4.2a.

To establish the biological significance of these motifs, we mutagenized motif A to the sequence 5′AAGCGCAAG and motif B to the sequence 5′GTGTGTG. Mutagenesis of either of these motifs individually in the promoter of the *UPRT* gene showed no significant change in promoter activity. Mutagenesis of both motifs within the *UPRT* promoter resulted in a seven-fold increase in reporter gene-expression, indicating that the two motifs possess redundancy in function to repress transcription (Figure 4.2b).

Genes encoding micronemal proteins

Micronemes are secretory organelles found in apicomplexan parasites and serve as compartments for the storage and trafficking of microneme proteins-a family of proteins that functions as ligands for host-cell receptors (Soldati et al. 2001). These proteins play a very important role in the active process of host-cell adhesion and invasion during the parasite life-cycle. We analyzed the upstream sequences of twelve microneme protein-encoding genes in *T. gondii* and corresponding upstream sequences for four orthologs in *E. tenella*. We identified two well-conserved sequence motifs in this data set that we subsequently selected for further

experimental characterization. Motif A is an 8 bp motif represented by the consensus sequence 5′GCGTCDCW (Figure 4.3a). It is found at least twice in the majority of the upstream regions, and is independent of orientation and position. This motif was also found upstream of *E. tenella* microneme genes. In the reverse orientation, this motif closely resembles the 5′WGAGACG motif that has been identified in previous studies to function as a regulatory element in several promoters of *T. gondii* (Soldati and Boothroyd 1995). Motif B is an 8 bp motif with a very well conserved sequence 5′SMTGCAGY (Figure 4.3a), the core "TGCA" nucleotides are conserved in 100% of the occurrences. This motif occurs once upstream of all the eleven microneme-encoding genes in *T. gondii*, but was not found in the corresponding orthologs in *E. tenella*. It is independent of position from the translational start site, and is always found on the forward strand (Figure 4.3a).

To characterize the functional significance of these conserved motifs, each of them was mutagenized to an 8 bp polyA sequence 5′AAAAAAAA. The mutagenesis of motif A in the *Mic8* promoter led to a ten-fold reduction in reporter activity, and the mutagenesis of motif B led to a three-fold reduction in reporter activity. When both motifs A and B were mutagenized in the same promoter, it led to an almost insignificant level of reporter expression (20-fold reduction in expression) (Figure 4.3b). From this data, we show that both motif A and motif B act as positive signals for gene-expression in the *mic 8* promoter, and together exert an additive effect on downstream gene-expression, as is evidenced from the low reporter read-outs when both A and B are mutagenized.

Ribosomal protein encoding genes:

Examination of stage-specific EST libraries indicates that the coccidia regulate *de novo* ribosome biosynthesis at the transcriptional level (Schaap et al. 2005). In a recent study, (Van Poppel et al.

2006) the authors examined a large set of cytoplasmic ribosomal proteins of *T. gondii* (79 genes in all) and describe the presence of two well-conserved motifs TRP-1: 5′TCGGCTTATATTCGG and TRP-2: 5′YGCATGCR identified by MEME in all the promoters. The sequence of TRP-2 is similar to the 8 bp element 5′TGCATGCA reported to be overrepresented in the non-coding regions of the apicomplexans *C. parvum*, *T. gondii* and *E. tenella* (Bankier et al. 2003). In a study of the promoter strengths of eight of the ribosomal protein genes, no correlation could be found between multiple occurrences of one or both motifs and promoter strength in the 8 promoters (Van Poppel, 2005). We conducted analyses on a subset of these genes (8 promoters) and recovered the motifs TRP-1 and TRP-2 as described by Van Poppel et al. (Figure 4.4a). We included these motifs in our analyses to ascertain if they functioned in a sequence-specific manner to affect promoter activity.

Motif TRP-1 in the *RPL9* promoter was mutagenized to the sequence 5′CGAAGTATGCCGAG, motif TRP-2 in the *RPL9* promoter was mutagenized to the sequence 5′TAAATAAA. Reporter assays revealed that the promoter containing A did not show any significant difference when compared to WT promoter activity. Mutagenesis of motif TRP-2 resulted in a two-fold reduction in promoter activity (Figure 4.4b).

Promoter organization in *T. gondii* has been studied in a few genes thus far (Soldati and Boothroyd 1995; Mercier et al. 1996; Ma et al. 2004; Kibe et al. 2005). In these studies, it has been observed that a gene-proximal region is necessary for minimal gene-expression and additional upstream sequence helps to enhance expression from the same promoter. However, very little is known about the mechanism of gene-regulation and the prevalence and type of transcriptional signals and regulatory apparatus in this organism. Analyses of genome sequences and individual gene-specific experiments point out to two deviations from what has been

observed in other model eukaryotes. One, canonical eukaryotic promoter elements such as the TATA box have not been found in *T. gondii* promoter regions (Soldati and Boothroyd 1995). Two, there is a stark paucity of known encoded specialized transcription factors encoded in the genome (Meissner and Soldati 2005). A similar scenario is seen in the other apicomplexan parasites *P. falciparum* and *C. parvum* (Gardner et al. 2002; Abrahamsen et al. 2004). This paradox can be explained by two ways: 1) These organisms do not employ a specialized transcriptional apparatus to regulate their genes or 2) A specialized transcriptional machinery exists, but is so divergent from known eukaryotic counterparts, that they cannot be detected by simple similarity-based searches. Recent studies have shown that the *T. gondii* genome encodes for a rich repertoire of histone-modifying enzymes, and epigenetic regulation has been purported to be responsible for stage-switiching in the parasite (Saksouk et al. 2005; Sullivan and Hakimi 2006). It is likely that histone-mediated regulation is responsible for regulation of genes to sizeable extent in *T. gondii*. Serial Analysis of Gene Expression (SAGE) of genes expressed during key life-cycle stages (Radke et al. 2005) have shown that the mRNA pool of *T. gondii* is highly dynamic and gene-expression is controlled in a time and stage-dependent manner. These studies have also shown that co-expressed genes in *T. gondii* do not cluster in the genome in terms of chromosomal location. Conversely, it has been seen in at least one case (*Eno1* and *Eno2*) that tandemly arranged genes are transcribed and expressed in a stage-specific manner (Kibe et al. 2005) and depend upon adjacent *cis*-regulatory elements. This implies that chromatin-mediated gene regulation cannot account for all of the regulation in the parasite, and other gene-specific controls of gene regulation must be involved in the control of gene expression. Indeed, searches of the *Plasmodium* genome sequence for transcription factors using secondary structure similarity have revealed the presence of putative transcription factors that

were missed in simple sequence-based searches (Callebaut et al. 2005). A divergent putative specialized transcription factor ApiAP2 has also been reported in the apicomplexa (Balaji et al. 2005). A large percentage of proteins in *T. gondii* are "hypothetical proteins" with no known function and might possibly encode parasite-specific functions including transcriptional regulatory proteins.

It is plausible that such highly divergent regulatory proteins utilize very different the *cis*-elements for their recruitment which would explain the absence of canonical *cis*-elements in the promoters studied thus far. Recently, two studies have reported the presence of novel *cis*-regulatory elements in promoters of genes regulated in a stage-specific manner (Ma et al. 2004; Kibe et al. 2005). The study of stage-specific gene regulation has been an active area of investigation, and small-scale microarray studies have also reported genes that are preferentially expressed in either stage (Cleary et al. 2002; Matrajt et al. 2002). However, gene regulation within the tachyzoite stage has not been well studied. The lack of a synchronized life-cycle of the parasites in *in vitro* culture makes it difficult to address gene regulatory questions in the actively multiplying tachyzoite, as any given population of cells in culture represent parasites at different points of their life-cycle.

We have exploited the availability of genome sequence for *T. gondii* to identify conserved upstream sequences in diverse groups of functionally related genes. We identified select over-represented motifs and tested their function *in vitro* in the parasite by specifically mutagenizing them and measuring reporter activity. For each group two candidate motifs were selected and characterized for their function in their endogenous promoter. We find that five out of eight motifs identified by *de novo* pattern finding show a significant role in promoter activity. Some of these motifs are evolutionarily conserved in *E. tenella* (Table 4.1). No similarities in these

motifs were seen for similar functional groups of genes studied in *C. parvum* (Mullapudi et al. 2007) a basal apicomplexan parasite. We have shown that conserved over-represented motifs play a definite role in gene-expression, and can affect promoter activity either positively or negatively (Figure 4.1-4.4). We selected motifs that were conserved in orthologs from the coccidian *E. tenella* for identifying evolutionarily conserved motifs possibly indicating a biological function. It is exciting that some elements that affected gene expression also exhibited cross-species conservation.

Our studies have shown that in spite of the lack of *a priori* knowledge of the nature of regulatory sequences or expression information, it is possible to identify putative *cis*-regulatory elements. We have shown for some computationally identified elements to be functionally relevant for gene expression. We have previously characterized putative *cis*-regulatory elements in the genome of *C. parvum* in a similar fashion where we established a correlation between over-represented elements and co-ordinate gene expression (Mullapudi et al, 2007). Among the functional groups common to both studies, we do not detect identical motifs in *T. gondii* and *C. parvum*. Given the evolutionary divergence, difference in genome organization and content and life-cycles between these two parasites, it is possible that they do not share some components of regulatory machinery.

Among the results described in this study, motif B identified in the glycolytic genes shares a good degree of sequence similarity to motif B identified in the microneme-coding genes (Figure 4.1a and Figure 4.3a). However, these motifs are not identical, and do not affect reporter expression the same way when mutagenized in their native promoters (Figure 4.1b and Figure 4.3b).

The use of an asynchronous population of parasites, as is the case here, is expected to dilute out effects to some extent, and this is reflected in the occasional small but significant change in promoter activity upon mutagenesis of a motif (Figure 4.1a, motif A; $p < 0.005$). We also realize that having selected promoters containing only one copy of the motif in order to assay the effect of the motif, we might have used intrinsically weak promoters and thus may not be able to see an effect of large magnitude. In spite of these limitations, our study has successfully identified at least five novel *cis*-regulatory elements and established the functional significance of one previously reported conserved upstream element (Figures 4.1-4.4). We have shown the applicability of computational techniques in a system where little is known about gene regulation and established the role of conserved upstream elements in gene expression in *T. gondii*. Thus *de novo* computational approaches serve as a great starting point to investigate biologically relevant questions in compact genomes when sequence is available. As microarray data become available for *T. gondii* (toxodb.org) it will be useful to identify genes that are co-regulated with the genes in this study and mine for the presence of similar motifs in additional genes.

Acknowledgements: We would like to thank Dr. Michael White and Michael Behnke for the *T. gondii* adapted luciferase vectors and related protocols, and Elizabeth Mathis and Karen Hermetz for technical assistance.

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G et al. (2004) Complete Genome Sequence of the Apicomplexan, *Cryptosporidium parvum*. Science 304: 441-445.

Alexander J, Scharton-Kersten TM, Yap G, Roberts CW, Liew FY et al. (1997) Mechanisms of innate resistance to *Toxoplasma gondii* infection. Philos Trans R Soc Lond B Biol Sci 352(1359): 1355-1359.

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.

Balaji S, Babu MM, Iyer LM, Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res 33(13): 3994-4006.

Bankier AT, Spriggs HF, Fartmann B, Konfortov BA, Madera M et al. (2003) Integrated mapping, chromosomal sequencing and sequence analysis of *Cryptosporidium parvum*. Genome Res 13(8): 1787-1799.

Berens RL, C. KE, Marr JJ (1995) Purine and pyridmidine metabolism. In: Marr JJ, Muller M, editors. biochemistry and molecular biology of parasites. London: Academic Press. pp. 89-117.

Black MW, Boothroyd JC (2000) Lytic cycle of *Toxoplasma gondii*. Microbiol Mol Biol Rev 64(3): 607-623.

Callebaut I, Prat K, Meurice E, Mornon JP, Tomavo S (2005) Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eukaryotes. BMC Genomics 6: 100.

Cleary MD, Singh U, Blader IJ, Brewer JL, Boothroyd JC (2002) *Toxoplasma gondii* asexual development: identification of developmentally regulated genes and distinct patterns of gene expression. Eukaryotic cell 1(3): 329-340.

Denton H, Roberts CW, Alexander J, Thong KW, Coombs GH (1996) Enzymes of energy metabolism in the bradyzoites and tachyzoites of *Toxoplasma gondii*. FEMS microbiology letters 137(1): 103-108.

Dzierszinski F, Mortuaire M, Dendouga N, Popescu O, Tomavo S (2001) Differential expression of two plant-like enolases with distinct enzymatic and antigenic properties during stage conversion of the protozoan parasite *Toxoplasma gondii*. Journal of molecular biology 309(5): 1017-1027.

Dzierszinski F, Popescu O, Toursel C, Slomianny C, Yahiaoui B et al. (1999) The protozoan parasite *Toxoplasma gondii* expresses two functional plant-like glycolytic enzymes. Implications for evolutionary origin of apicomplexans. J Biol Chem 274(35): 24888-24895.

Fox BA, Bzik DJ (2002) De novo pyrimidine biosynthesis is required for virulence of *Toxoplasma gondii*. Nature 415(6874): 926-929.

Gardner MJ, Hall N, Fung E, White O, Berriman M et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419(6906): 498-511.

Gilbert RE, Gras L, Wallon M, Peyron F, Ades AE et al. (2001) Effect of prenatal treatment on mother to child transmission of *Toxoplasma gondii*: retrospective cohort study of 554 mother-child pairs in Lyon, France. Int J Epidemiol 30(6): 1303-1308.

Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. Journal of molecular biology 296(5): 1205-1214.

Kibe MK, Coppin A, Dendouga N, Oria G, Meurice E et al. (2005) Transcriptional regulation of two stage-specifically expressed genes in the protozoan parasite *Toxoplasma gondii*. Nucleic Acids Res 33(5): 1722-1736.

Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS (2003) ToxoDB: accessing the *Toxoplasma gondii* genome. Nucleic Acids Res 31(1): 234-236.

Ma YF, Zhang Y, Kim K, Weiss LM (2004) Identification and characterisation of a regulatory region in the *Toxoplasma gondii* hsp70 genomic locus. Int J Parasitol 34(3): 333-346.

Manger ID, Hehl A, Parmley S, Sibley LD, Marra M et al. (1998) Expressed sequence tag analysis of the bradyzoite stage of *Toxoplasma gondii*: identification of developmentally regulated genes. Infect Immun 66(4): 1632-1637.

Matrajt M, Donald RG, Singh U, Roos DS (2002) Identification and characterization of differentiation mutants in the protozoan parasite *Toxoplasma gondii*. Mol Microbiol 44(3): 735-747.

Mattsson JG, Erhardt H, Soldati D (1997) In control of its fate: gene regulation in *Toxoplasma gondii*. Behring Inst Mitt(99): 25-33.

Meissner M, Soldati D (2005) The transcription machinery and the molecular toolbox to control gene expression in *Toxoplasma gondii* and other protozoan parasites. Microbes Infect 7(13): 1376-1384.

Mercier C, Lefebvre-Van Hende S, Garber GE, Lecordier L, Capron A et al. (1996) Common cis-acting elements critical for the expression of several genes of *Toxoplasma gondii*. Mol Microbiol 21(2): 421-428.

Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS et al. (2005) The transcriptome of *Toxoplasma gondii*. BMC Biol 3: 26.

Saksouk N, Bhatti MM, Kieffer S, Smith AT, Musset K et al. (2005) Histone-modifying complexes regulate gene expression pertinent to the differentiation of the protozoan parasite *Toxoplasma gondii*. Molecular and cellular biology 25(23): 10301-10314.

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning A Laboratory Manual. Cold Spring Harbor.

Schaap D, Arts G, van Poppel NF, Vermeulen AN (2005) De novo ribosome biosynthesis is transcriptionally regulated in *Eimeria tenella*, dependent on its life-cycle stage. Mol Biochem Parasitol 139(2): 239-248.

Soldati D, Boothroyd JC (1993) Transient transfection and expression in the obligate intracellular parasite *Toxoplasma gondii*. Science (New York, NY 260(5106): 349-352.

Soldati D, Boothroyd JC (1995) A selector of transcription initiation in the protozoan parasite *Toxoplasma gondii*. Mol Cell Biol 15(1): 87-93.

Soldati D, Dubremetz JF, Lebrun M (2001) Microneme proteins: structural and functional requirements to promote adhesion and invasion by the apicomplexan parasite *Toxoplasma gondii*. International journal for parasitology 31(12): 1293-1302.

Striepen B, Pruijssers AJP, Huang J, Li C, Gubbels MJ et al. (2004) Gene transfer in the evolution of parasite nucleotide biosynthesis. Proceedings of the National Academy of Sciences, USA 101(9): 3154-3159.

Struhl K (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell 98(1): 1-4.

Sullivan WJ, Jr., Hakimi MA (2006) Histone mediated gene activation in *Toxoplasma gondii*. Mol Biochem Parasitol 148(2): 109-116.

Van Poppel N, Welagen J, Vermeulen AN, Schaap D (2006) The complete set of *Toxoplasma gondii* ribosomal protein genes contains two conserved promoter elements. Parasitology 133(Pt 1): 19-31.

Yang S, Parmley SF (1995) A bradyzoite stage-specifically expressed gene of *Toxoplasma gondii* encodes a polypeptide homologous to lactate dehydrogenase. Mol Biochem Parasitol 73(1-2): 291-294.

Table 4.1 List of genes used in this study

A list of genes and lengths of their upstream regions that were used to identify regulatory motifs. (*) indicates that the corresponding orthologs in *E. tenella* was obtained and added to the search

Table 4.1

GENE SYMBOL	NAME	LENGTH
GLYCOLYSIS		
*HK**	Hexokinase	1000
*G6PI**	Glucose-6-P-Isomerase	2000
PFK	Phosphofructokinase	2000
*ALD**	Aldolase	1370
*TPI**	Triose-P-Isomerase	2000
*GAPDH**	Glyceraldehydye-3-Phosphate Dehydrogenase	2000
PGK	Phosphoglycerate kinase	2000
PGM	Phosphoglucomutase	1800
*ENO**	Enolase	2000
PyK	Pyruvate Kinase	
NUCLEOTIDE METABOLISM		
*AK**	Adenosine Kinase	2000
*CTPS**	Ctydine Synthase	2000
*DCDA**	Deoxycyrdine Deaminase	2000
DHFR-TS	Dihydrofolate reducatase-Thymidine synthase	2000
*GMPS**	Guanidine Monophosphate Synthase	2000
RDPR	Ribonucelotide Diphosphate Reductase	2000
UPRT	Uracil Phosphoribosyl Transferase	2000
AT	Adenosine Transferase	2000
		2000
MICRONEMAL PROTEIN GENES		
*MIC1**	Microneme 1	1500
*MIC2**	Microneme 2	1500
MIC3	Microneme 3	1500
*MIC4**	Microneme 4	1500
*MIC5**	Microneme 5	1500
*MIC6**	Microneme 6	1500
MIC7	Microneme 7	1500
MIC8	Microneme 8	1500
MIC9	Microneme 9	1500
*MIC10**	Microneme 10	1500
MIC11	Microneme 11	1500
M2AP	Microneme-2-associated protein	1500
RIBOSOMAL PROTEIN GENES		
RPS29	Ribosomal protein S29	800
RPS38	Ribosomal protein S38	1000
RPS3	Ribosomal protein S3	1000
RPS13	Ribosomal protein S13	1000
RPL9	Ribosomal protein L9	1200
RPS25	Ribosomal protein S25	1300
RPS10	Ribosomal protein S10	700
RPL25	Ribosomal protein L25	1000

Table 4.2 List of primers

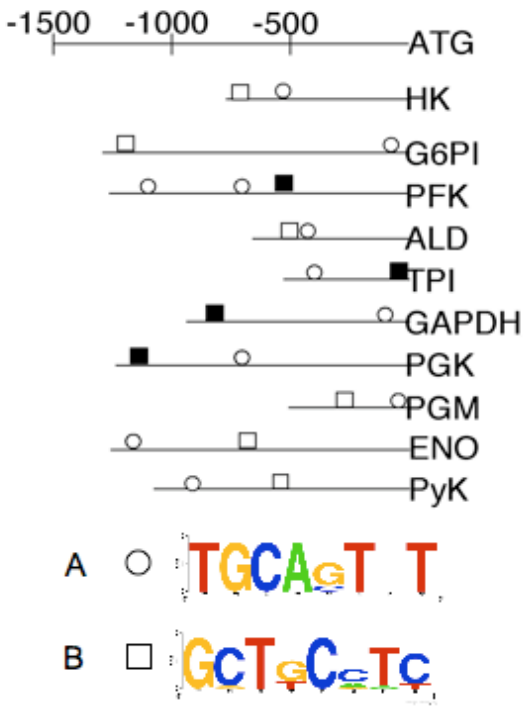
Forward and reverse primers used for PCR-amplification of group-specific promoters from *T. gondii* genomic DNA. 'F' denotes forward primer, 'R' denotes reverse primer. All PCR reactions were carried out at a 95°C denaturation, 60°C annealing and 72°C extension temperature for 30 cycles.

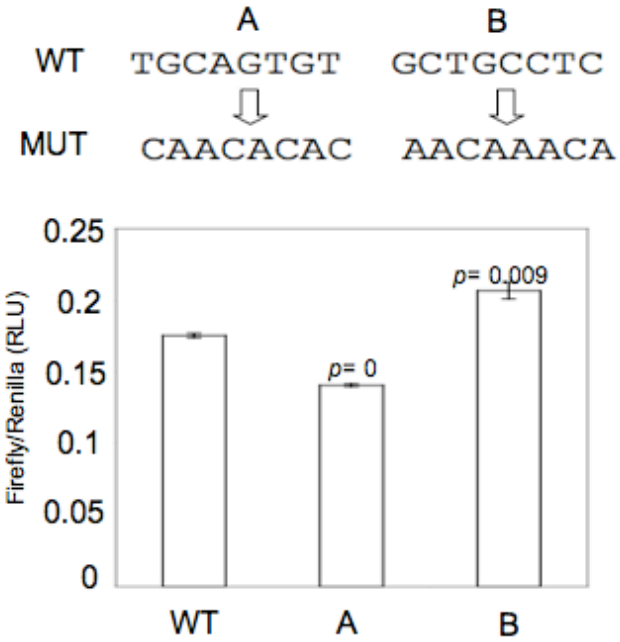
Table 4.2

NAME	SEQUENCE
ENO2F	5'GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGTGAAAACACCGAAGAAAGG
ENO2R	5'GGGGACCACTTTGTACAAGAAAGCTGGGTCCGCCATTTGTTGGAAGTTGAGTAACGG
MIC8F	5'GGGGACAAGTTTGTACAAAAAAGCAGGCTTCACACGACAAATATATGGCC
MIC8R	5'GGGGACCACTTTGTACAAGAAAGCTGGGTCCGCCATTCTGTCTTGAGGAACGA
UPRTF	5'GGGGACCACTTTGTACAAGAAAGCTGGGTCCGCCATTTTAGAAGCCCTGTGGAAAG
UPRTR	5'GGGGACCACTTTGTACAAGAAAGCTGGGTCCGCCATTTTAGAAGCCCTGTG
RPL9F	5'GGGGACAAGTTTGTACAAAAAAGCAGGCTTCAGAGGAAGAGCCGTCAGAAC
RPL9R	5'GGGGACCACTTTGTACAAGAAAGCTGGGTCCGCCATGCGGGAGAGGGGAGAGTC

Figure 4.1-4.4: Candidate motifs identified in each group of genes, mutagenesis and results of reporter assays: Panel (a) shows the positions of the motifs with respect to the translational start and sequence logos representing the consensus sequence for each motif. The gene names are abbreviated as shown in Table 4.1. Motif A found in both *E. tenella* and *T. gondii* is denoted by a circle and motif B exclusive *to T. gondii* is denoted by a square. Filled in shapes denote motifs on the opposite strand.

The WT motifs and their mutagenized versions are represented at the top of panel (b). The graphs depict luciferase activity in ratios of firefly/renilla activity in relative luciferase units (RLU) from the different constructs, containing either WT or mutagenized versions of A, B, or both motifs. All luciferase readings are relative to an internal control (α-tubulin-promoter-renilla-luciferase)

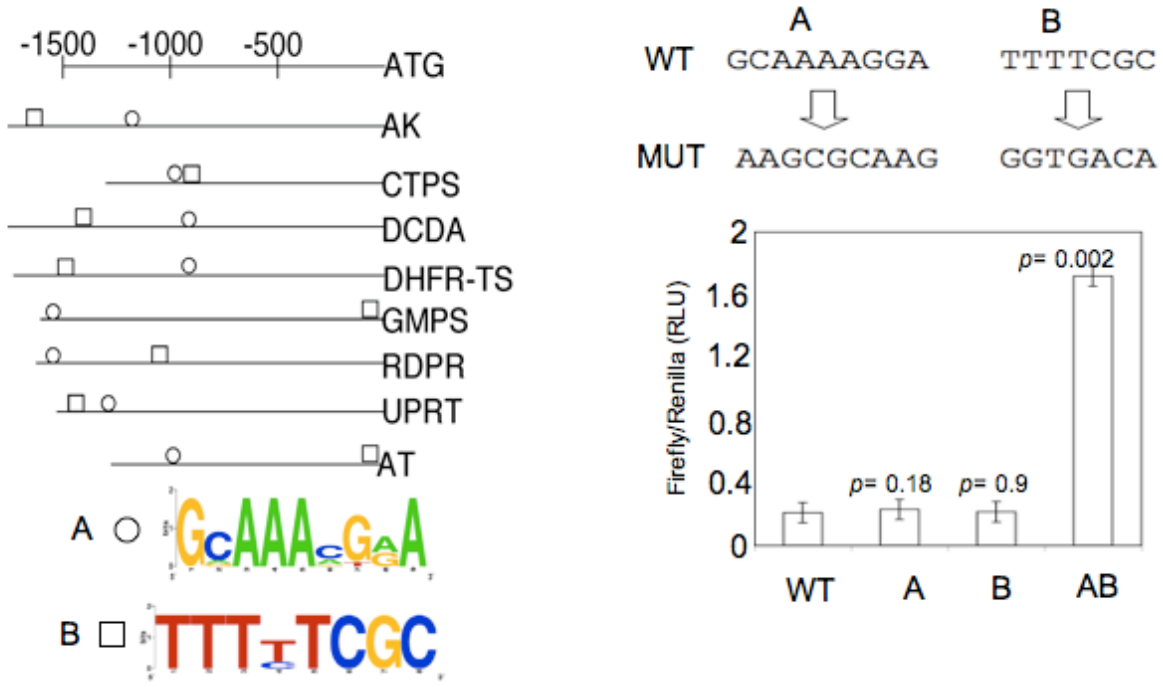




(a) **(b)**

Figure 4.1

(a) Candidate upstream motifs for the glycolytic enzymes and (b) results of reporter assays with the *Eno2* promoter



(a) **(b)**

Figure 4.2

(a) Candidate upstream motifs for the nucleotide metabolism enzymes and (b) results of reporter assays with the *UPRT* promoter

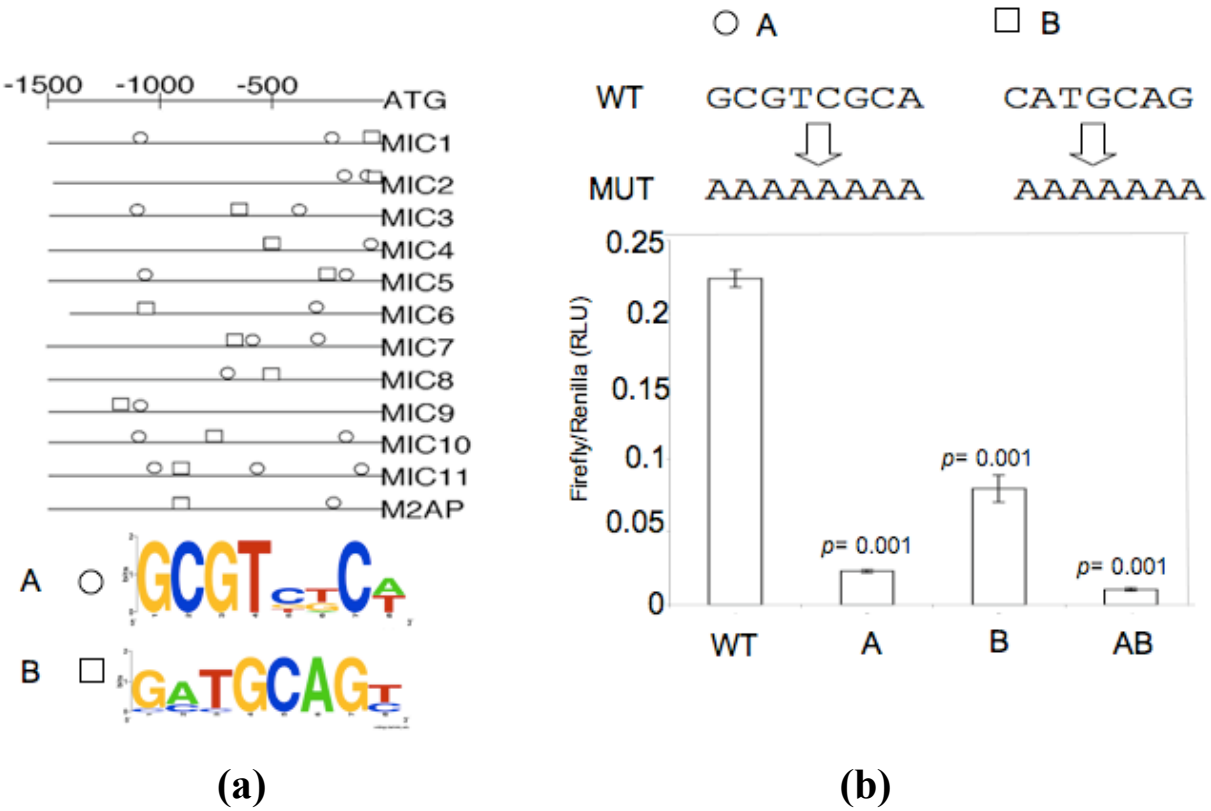


Figure 4.3

(a) Candidate upstream motifs for the micronemal proteins and (b) results of reporter assays with the *Mic8* promoter

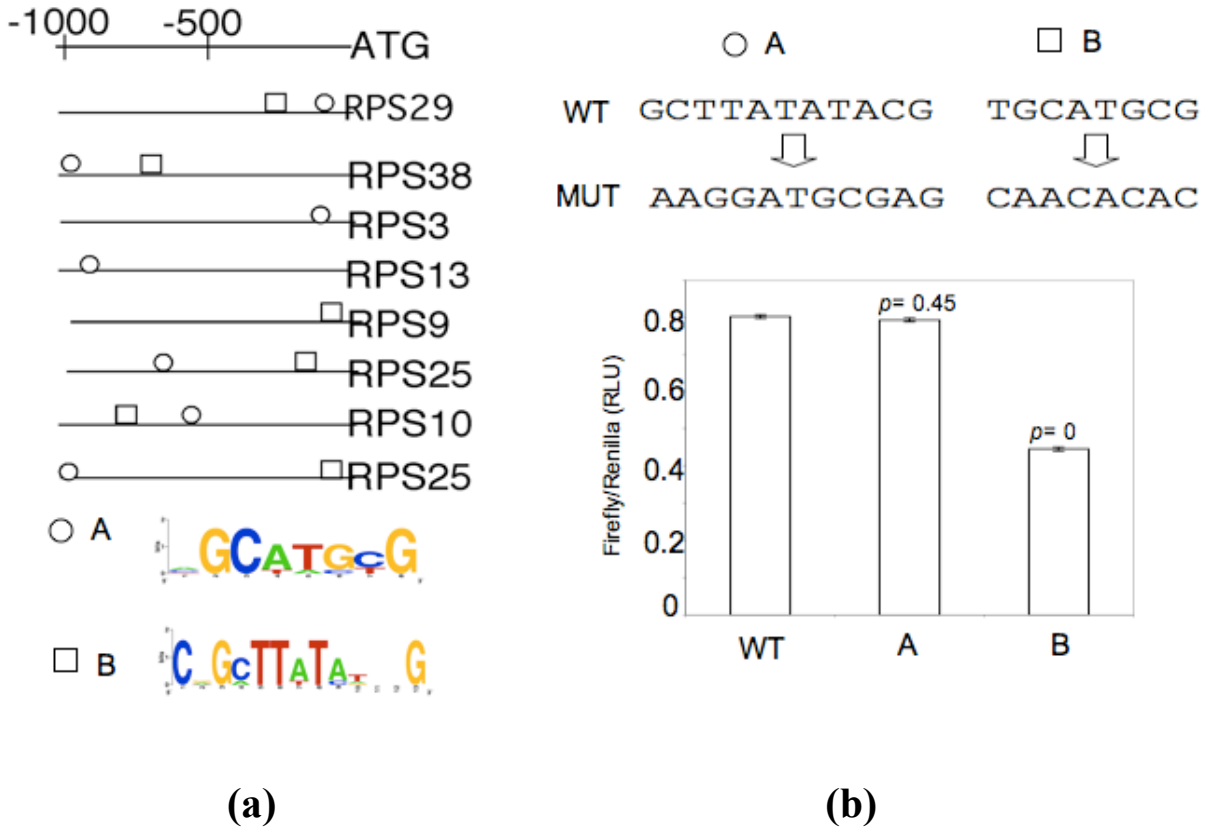


Figure 4.4

(a) Candidate upstream motifs for the ribosomal proteins and (b) results of reporter assays with the *RPL9* promoter

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

5.1 Conclusions

5.1.1 Groups of functionally-related genes share conserved upstream elements

We have mined genome sequence data to detect over-represented conserved upstream elements within distinct groups of functionally related genes in *Cryptosporidium parvum* and *Toxoplasma gondii*. Our work thus far has shown that conserved upstream elements show a clear association with co-expressed genes in *C. parvum* (Chapter 3) and influence downstream gene-expression in *T. gondii* (Chapter 4). It is seen that different groups of functionally related genes share different *cis*-elements. The elements identified in this study do not bear resemblance to previously identified eukaryotic *cis*-regulatory elements. Our results indicate that transcriptional regulation via *cis*-regulatory elements is a significant component of gene-regulation in *T. gondii* and is not only restricted to developmentally-regulated genes. The apparent lack of identifiable specialized transcription factors in the apicomplexan genomes can be explained in part by the presence of divergent transcriptional machinery that utilizes correspondingly divergent regulatory signals like those identified in this study. Reporter assays in *T. gondii* indicate both positive and negative functions for the various elements identified in regulating downstream gene expression.

5.1.2 *De novo* computational methods can be successfully applied to yield biologically significant results

Typically, computational identification of *cis*-regulatory elements is aided by prior knowledge of gene-expression profiles or the sequence and/or organization of regulatory

elements. In the absence of any of these types of information, we conducted *de novo* searches to identify novel *cis*-regulatory elements in several diverse groups of putatively co-regulated genes. We established a rule-based method to filter the output of pattern finding programs and shortlist candidate elements for experimental validation. Our work is among the first few reports on the identification of functionally significant elements without any *a priori* knowledge (Cohen et al. 2006). The initial identification of novel *cis*-regulatory elements by this method serves as an excellent starting point for experimental validation and also provides an alternative approach to the detection of biologically significant motifs in promoters without carrying out more traditional sequential promoter bashing and deletion experiments.

5.2 Future Directions

5.2.1 Identification of putative transcription factors in the apicomplexa

A logical next step based on our findings would be to investigate the presence of proteins from parasite nuclear lysates that specifically bind the candidate motifs by way of gel-shift assays. Gel-shift assays or EMSAs (Electrophoretic Mobility Shift Assay) test the ability of a short, labeled DNA sequence (30 – 300bp) to form DNA-protein complexes with specific proteins from either a mixuture of proteins (nuclear lysate) or in the presence of purified proteins. The specificity of this interaction can be tested by using mutagenized versions of the DNA sequence to test for abolished binding, and excess unlabeled DNA fragments to out-compete the labeled fragment for binding to the protein. It would be important to control for any possible host-cell contamination of the parasite nuclear lysates by carrying out similar assays with host-cell nuclear extract. In the case of *C. parvum*, this is an experimentally challenging step, given the difficulty of purifying the parasites from the host cell in a sufficient quantity to obtain a concentrated nuclear extract. Proteins of interest, such as transcription factors, are

expected to be present in very small quantities in nuclear extracts, making it difficult to detect or demonstrate specific binding activity from nuclear lysates from a heterogenous population of parasites. However, there have been recent reports of improved culture methods to yield large numbers of parasites (Steve Upton, personal communication). These improvements should prove useful to follow up on our studies and further investigate *C. parvum* gene-regulation in the context of detection of specific proteins.

In *T. gondii*, there have been recent reports of EMSAs with both tachyzoite and bradyzoite nuclear extracts, leading to the discovery of specific binding activity of proteins in the nuclear extract for specific DNA elements (Ma et al. 2004; Kibe et al. 2005; Cohen et al. 2006). These experiments have benefited from the use of regulatory motifs pertaining to stage-specific genes and corresponding parasite populations to obtain nuclear lysates, sufficient for the enrichment of specific proteins in the lysates. We have tried to implement these methods to investigate the functions of *cis*-regulatory motifs discovered in our study (Chapter 4), but have met with varying degrees of success in terms of the reproducibility of results. (Figure A1). The *cis*-regulatory motifs identified in chapter 4 are not specific for a certain developmental stage in the parasite life-cycle. In the absence of a synchronized cell-population and specific discriminatory conditions to enrich the parasite population for putative binding proteins corresponding to the regulatory elements identified, the detection of such activity has proven difficult. Once a reproducible, specific binding activity can be ascertained, there are several methods to further characterize the protein of interest, including the use of mass spectrophotometric methods to identify the constituents of the extract binding to a particular motif and tandem affinity purification methods to isolate the DNA-binding protein from a mixture of proteins (Deng et al. 2003; Sterling et al. 2006).

5.2.2 Correlation and extension of our work with large-scale expression analyses

In addition to the functional studies described above for the identification and characterization of *cis*-regulatory elements and their cognate transcription factors, the recent explosion in whole-genome sequence data opens up many possibilities to study gene regulation on a genome-wide scale. Whole genome sequences provide the starting material for large-scale expression analysis via microarray studies; for comparative sequence analysis and for the application of sophisticated search methods to annotate orphan proteins with transcription-related function if they are found. Additionally, the detection of *cis*-regulatory elements can be vastly improved and algorithms can be trained after a small number of sites are verified to have function. Similar studies in yeast and in the humans have led to the creation of databases that serve as a repository of regulatory-sequence information and allow for the detection of new regulatory clusters and networks (Perier et al. 2000; Praz et al. 2002).

Now is indeed an exciting time in the field of apicomplexan gene regulation. At the time the studies described in this thesis were begun, there was very sparse literature and experimental evidence on gene-regulatory methods in apicomplexan parasites. The availability of whole-genome sequences and the development of several new molecular tools (Meissner et al. 2007) to validate computational predictions together serve as a very powerful combination to gain better insight into mechanistic questions about gene regulation.Currently, whole genome sequences are available for 5 species of *Plasmodium*, two *Cryptosporidium* species, two *Theileria* spp. and for *T. gondii* {Abrahamsen, 2004 #2447; Gardner, 2005 #2967; Gardner, 2002 #681; Xu, 2004 #2643; Pain, 2005 #2968; Pain, 2005 #3331; Aurrecoechea, 2007 #3294}. Other apicomplexan genomes close to completion include *Eimeria tenella*, *Babesia bovis* , *Neospora caninum* and *C. muris* (http://www.sanger.ac.uk/Projects/Protozoa/). Most of the sequence information is being

made available at a single website (http://apidb.org, Aurrecoechea et al. 2007) which facilitates cross-species comparisons and bulk sequence retrieval. In chapter 3 and 4, I have described the development of in-house computational tools to analyze whole-genome information (PERL scripts, custom-built databases, etc.). A common relational database with a variety of options for retrieval and visualization of inter-coding sequences and orthologs will obviate the need for such custom tools to some extent, enabling more researchers to ask relevant questions and exploit this sequence information with relative ease.

In *C. parvum*, experiments are underway to characterize comparative RT-PCR based expression profiles for all the 3396 predicted genes (Mitch Abrahamsen & Jessica C. Kissinger, personal communication) as was described earlier (Chapter 3). Once these expression profiles are available, genes can be clustered into larger groups based on their expression profiles and their upstream regions can be analyzed for the presence of *cis*-element candidates reported here, and for the discovery of additional novel *cis*-elements. Conversely, the discriminatory power of the reported motifs can be tested by recovering other genes in the genome that contain these upstream motifs and correlating their expression profiles with the group of genes used in this study. These ideas provide ample scope to reiterate and improve upon *cis*-element modeling and prediction in *C. parvum*.

In *T. gondii*, microarrays have recently been developed to study gene-expression on a global scale. Thanks to concerted efforts by the *Toxoplasma* research community, photolithographic chips covering a majority of coding sequences in *T. gondii* have been designed and are soon to be made available at a very low cost (http://toxodb.org). These chips contain probes to all predicted nuclear and organellar genes of *T. gondii* at a coverage of 11 probes per gene. In addition to this, they contain probes to non-coding sequences (5′ UTRs) to ~5000 genes

and intergenic and intronic sequences to all the genes on chromosome Ib. These chips can thus be used to analyze genome-scale gene expression in *T. gondii* and can provide insight into variations in expression profiles in a time- and stage-dependent fashion. They can provide better resolution than that obtained from the SAGE data since we are now in a position to profile gene expression at a finer scale; for example the expression of genes within the tachyzoite stage alone can be analyzed over smaller time-points. This would tell us about the regulation of genes involved in host-cell invasion, cell division and other processes. However, these analyses are possible only if a synchronized cell population is available.

A second useful application of whole genome arrays is to conduct Chromatin immunoprecipitation experiments (ChIP on chip) to identify the regions of the genome to which specific proteins are preferentially bound. The basic premise of this method involves cross-linking DNA-protein complexes in a cell, shearing the chromatin and using specific antibodies against proteins of interest to immunoprecipitate the protein attached to its target DNA. The associated DNA can then be identified by using it to probe a genomic DNA microarray and thus identify the target regions of the protein in the genome. (Lee et al. 2006). The most common application is to use antiodies to modified histones to analyze regions of the genome that are preferentially associated with particular modified histones such as methylated histones, acetylated histones, etc. This assay can be well quantitated and provide data regarding the enrichment of specific proteins at certain sites on the genome. In the Affymetrix arrays described above, the whole genome intergenic and intronic regions for chromosome Ib in *T. gondii* have been spotted on the array, to allow for identification and characterization of promoter regions that are preferentially associated with specific types of histones. Given that chromatin-mediated gene regulation is expected to play an important role in apicomplexan gene-

regulation, these studies will provide important clues about the extent to which promoters are associated with modified histones and the consequent effect on the region in terms of gene-expression. Results of ChIP can then be combined with expression analyses to provide us with the correlation between chromatin modification and gene expression.

Genome-wide studies and functional genomics will provide useful insight into trends of gene expression and allow us to design specific experiments to investigate the expression of genes. High throughput methods will thus help to identify global trends of regulation in the apicomplexan parasites, and these can be followed up with more gene-specific studies to understand mechanistic principles underlying gene-regulation in the Apicomplexa. This knowledge can further be extended to elucidate higher levels of regulatory networks in apicomplexan parasites.

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G et al. (2004) Complete Genome Sequence of the Apicomplexan, *Cryptosporidium parvum*. Science (New York, NY 304: 441-445.

Aurrecoechea C, Heiges M, Wang H, Wang Z, Fischer S et al. (2007) ApiDB: Integrated Resources for the Apicomplexan Bioinformatics Resource Center Nucleic acids research 35: In Press.

Cohen CD, Klingenhoff A, Boucherot A, Nitsche A, Henger A et al. (2006) Comparative promoter analysis allows de novo identification of specialized cell junction-associated proteins. Proceedings of the National Academy of Sciences of the United States of America 103(15): 5682-5687.

Deng WG, Zhu Y, Montero A, Wu KK (2003) Quantitative analysis of binding of transcription factor complex to biotinylated DNA probe by a streptavidin-agarose pulldown assay. Analytical biochemistry 323(1): 12-18.

Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM et al. (2005) Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. Science (New York, NY 309(5731): 134-137.

Gardner MJ, Hall N, Fung E, White O, Berriman M et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419(6906): 498-511.

Kibe MK, Coppin A, Dendouga N, Oria G, Meurice E et al. (2005) Transcriptional regulation of two stage-specifically expressed genes in the protozoan parasite *Toxoplasma gondii*. Nucleic acids research 33(5): 1722-1736.

Ma YF, Zhang Y, Kim K, Weiss LM (2004) Identification and characterisation of a regulatory region in the *Toxoplasma gondii* hsp70 genomic locus. International journal for parasitology 34(3): 333-346.

Meissner M, Agop-Nersesian C, Sullivan WJ, Jr. (2007) Molecular tools for analysis of gene function in parasitic microorganisms. Appl Microbiol Biotechnol.

Pain A, Renauld H, Berriman M, Murphy L, Yeats CA et al. (2005a) Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. Science (New York, NY 309(5731): 131-133.

Perier RC, Praz V, Junier T, Bonnard C, Bucher P (2000) The eukaryotic promoter database (EPD). Nucleic acids research 28(1): 302-303.

Praz V, Perier R, Bonnard C, Bucher P (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic acids research 30(1): 322-324.

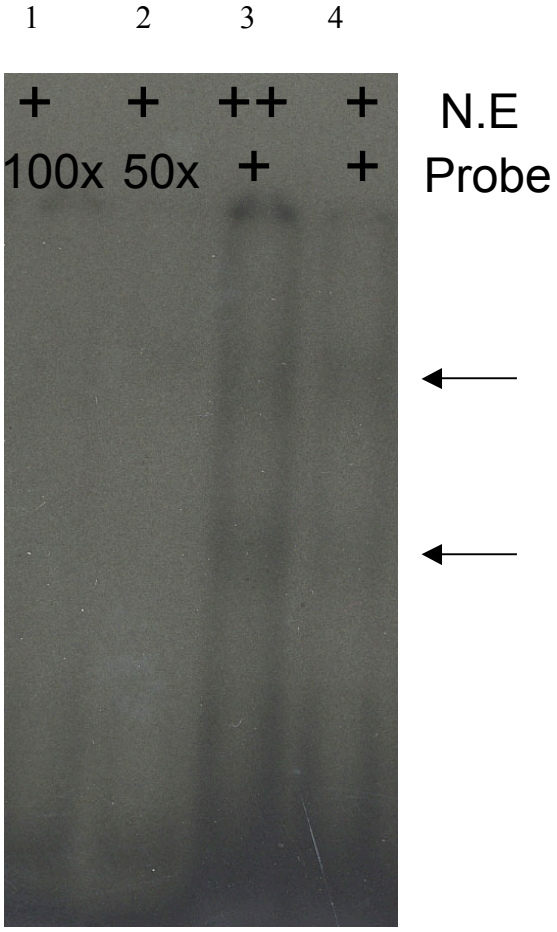
Sterling JA, Wu L, Banerji SS (2006) PARP regulates TGF-beta receptor type II expression in estrogen receptor-positive breast cancer cell lines. Anticancer research 26(3A): 1893-1901.

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM et al. (2004) The genome of *Cryptosporidium hominis*. Nature 431(7012): 1107-1112.

APPENDIX A1: EMSA with Mic8 motif B

Two shifted bands (denoted by arrows) were observed when the radiolabeled Mic8 probe was incubated with the *T. gondii* nuclear extract, increasing in intensity as the concentration of nuclear extract added increased (Lanes 3 and 4, indicated by + signs). Addition of 50x and 100x unlabeled probe to the probe-lysate mixture resulted in elimination of the shifted bands. (Lanes 1 and 2).

FIGURE A1



Materials and Methods (Figure A1)

Parasite culture: 8 –10 T150s were inoculated with *T. gondii* (RH strain) tachyzoites from a freshly lysed culture of hTERT cells. Tachyzoites were allowed to grow until they lysed hTERT cells spontaneously and were harvested by passing through a 27g needle and filtered twice through 3 μm nucleopore membrane filter.

Preparation of nuclear extract: Parasites harvested as above were then centrifuged at 1500g for 10 min. at 4°C and the pellet was washed twice in cold PBS (centrifuged at 1500g for 10 min. at 4°C between washes). The pellet was then resuspended in ice cold 500 ul Buffer B (20 mM HEPES pH 8, 100 mM KCl, 0.5 mM DTT, 0.2 mM EDTA, 20% glycerol and 1 tablet protease inhibitor cocktail (Roche) per 2 ml of buffer), centrifuged at 450g for 10 min at 4°C and then resuspended in 200 ml Buffer B. This was done in a 2 ml appendorf tube to make it easy to sonicate. Parasite suspension was sonicated for 20 seconds on ice (Branson sonicator (VWR), amplitude = 60%), with care taken to avoid bubbling.

The parasite suspension was then centrifuged at 3000g at 4°C and the supernatant was aliquoted in 30 μl volumes and stored at –80°C. The protein concentration in the nuclear lysate was determined by the Bradford method.

Electrophoresis mobility shift assays: Electrophoresis mobility shift assays (EMSA) were performed using a gel-shift assay kit (Promega) following the manufacturer's instructions. Briefly, DNA fragments were generated by PCR using primers (long probes) or by commercial synthesis (40 bp probes). Commercially synthesized probes were annealed to produce double-stranded oligomers which were then end-labeled with [γ-32P]-ATP using 10 U of polynucleotide kinase. The labelled DNA fragments or oligonucleotides were purified by passage through a SephadexG-50 column. The probes were used immediately or stored at –20°C and used within

one week. For gel shift assays, labelled DNA fragment or double-stranded oligonucleotides (20 fmol) were incubated for 30 min at 37°C with 15 - 25 μg of nuclear extracts in binding buffer (10 mM Tris–HCl at pH 7.5, 50 mM NaCl and 0.5 mM DTT) containing 1 μg Poly(dI–dC):Poly(dI–dC) (Amersham), 0.01% NP40 and 10% glycerol. For competition experiments, 50 to 100-fold excess of self or unrelated unlabelled DNA fragments were added to the binding reaction together with the nuclear extracts before addition of the labeled probe. The reactions were run on a 6% native polyacrylamide gel in 0.5x TBE running buffer. Best results were obtained when the gel was run at 4°C for 1.5 hours at 150 volts, after a pre-run using the same conditions.