GENE EXPRESSION IN THE TRYPANOSOMATID *LEISHMANIA*

by

JULIE DANGREMOND STANTON

(Under the direction of Kojo Mensa-Wilmot)

ABSTRACT

*Leishmania* species are the etiological agents of the disease leishmaniasis. Over 12 million people are infected with leishmaniasis, and ten percent of the world's population is at risk of contracting the disease. Trypanosomatids exert little control at the level of transcription because most promoters are active by default. As a result these organisms rely heavily on post-transcriptional mechanisms for regulating gene expression. The untranslated regions (UTRs) of mRNA may contribute to this type of control.

The purpose of this study was to investigate the effects of 5' UTRs on gene expression in *Leishmania*. We focused on the AUG-proximal region (APR), defined as the nucleotides at positions -3, -2, and -1 upstream of the translation initiation codon. Eight 12-nucleotide sequences differing in the APR were studied in the context of a 180-nucleotide *Leishmania* 5' UTR. Using stable transfectants of *Leishmania tropica*, up to a 6000-fold difference in reporter protein expression could be observed. Steady state levels of reporter mRNA did not reflect the dramatic variation in protein abundance, indicating that APRs act post-transcriptionally. To explore the possibility that APRs acted on translation, their effect on ribosome recruitment was examined. The protein expression from some APRs correlated with the extent of polysome formation on reporter mRNA. We conclude that APRs can affect translation in *Leishmania*.

To further study the role of the 5' UTR, tri-nucleotides in the APR of 200 *Leishmania major* genes were analyzed. Some tri-nucleotides were observed up to $7.724 \times 10^5$ times more than expected if selection was random. Twenty-five percent of the tri-nucleotides were absent from the data set, a 6-fold higher frequency than expected. We conclude that selection of tri-nucleotides in the APR of *Leishmania* 5' UTRs is not random. Using nearest neighbor analysis, patterns containing frequently used and rare tri-nucleotides were found. Together, our experimental and genomic data indicate that APRs have functional significance in *Leishmania*.

INDEX WORDS: *Leishmania*, Untranslated regions, Gene expression, Translation, Polysome, Ribosome, Nearest neighbor analysis
GENE EXPRESSION IN THE TRYPANOSOMATID LEISHMANIA

by

JULIE DANGREMOND STANTON

B.S., Truman State University 1996

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2003
© 2003

Julie Dangremond Stanton

All Rights Reserved
GENE EXPRESSION IN THE TRYPANOSOMATID *LEISHMANIA*

by

JULIE DANGREMOND STANTON

Major Professor: Kojo Mensa-Wilmot

Committee: Haini Cai
Claiborne Glover
James Lauderdale
Kelley Moremen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2003
DEDICATION

This dissertation is dedicated
to Dr. Diane Janick-Buckner
for inspiring me to become an educator and a scientist
and to my dear husband James
for supporting me in that great endeavor.
ACKNOWLEDGEMENTS

I am happy to have the opportunity to properly thank my major professor, Dr. Kojo Mensa-Wilmot, for all of his help and especially his support over the last six years. I truly believe that had I not found my way into his lab, I would not have finished this degree. When I came to work with him I had no research experience, and I am amazed at how far his guidance and training have taken me. Kojo is not only an excellent teacher but also a caring mentor who helped me during some challenging times. As I complete my Ph.D., I feel both thankful and blessed to have been his student.

I thank the members of my committee, Dr. Haini Cai, Dr. Claiborne Glover, Dr. James Lauderdale, Dr. Kelley Moremen, and Dr. Kathy Spindler for sharing their time and giving helpful suggestions. I appreciate the generosity of Dr. Judy Willis, Dr. Michael Terns, Dr. Claiborne Glover, and Dr. Mark Brown, who allowed me to use equipment in their labs. I thank Dr. Joe Crim, Dr. Marcus Fechheimer, and Dr. Michele Simpson, whose encouragement of my research, teaching, and service endeavors has meant so much to me. I also want to acknowledge Dr. Diane Janick-Buckner, Dr. Brent Buckner, Dr. Nancy Sanders, and Dr. Ann Moody of Truman State University for sharing their enthusiasm for science and dedication to teaching with me.

Past and present Mensa-Wilmot lab members have made working enjoyable, and I am thankful for all that I learned from them. They include Dr.
Meredith Teilhet Morris, Dr. Dora Armah, Kimberly Butler, Dr. Mohammed Rashid, Robyn Howard, Kameko Nichols, Peter Harri, Dr. Zhifeng Zheng, Frank Hardin, Bhargavi Patham, Sandesh Subramanya, Brad Allgood, Kitty Williams, and Dimple Bosu. Those who worked side by side with me in room 609 deserve to be acknowledged for making it such a fun place to be. I also appreciate Meredith for training me with patience and kindness, Dora for being an excellent role model, and Kimberly for her willingness to help without being asked. I am grateful for these friends and others I made during my time here.

I want to acknowledge my family for all they have given me. I thank my dad for sharing the joy of learning with me and for lovingly supporting me throughout my life. I thank my mom for encouraging me to work hard, persevere, and always try my best. I thank my sister Pam for being available to talk at any time and for making me smile when things got rough. I am also grateful to the Stanton family, who I enjoyed getting to know during my time in Georgia.

I completed this dissertation in parallel with my oldest friend Michelle Byrtek, who was finishing her Ph.D. in statistics. I am so thankful for her understanding as we shared in this unique situation, as we have in many other life events. I also appreciate Joanna Walczykiewicz and Traci Meziere, two friends who offered good advice and simply listened on many occasions.

Finally, I could not have reached this goal without my dearest friend and husband, James Stanton. Graduate school brought out the best and worst in me; it challenged me in ways I could never anticipate. James' unwavering
support and steadfast love gave me the courage to see this long journey to its completion. I drew strength from his encouragement and great comfort from his presence. As a result, I feel that in a way this is an accomplishment we made together, and not just a degree I earned alone. It is a joy and a blessing to share my life with him, and so I thank him for all of this and much much more.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS .................................................................................................................. v

CHAPTER

1 INTRODUCTION AND LITERATURE REVIEW ................................................................. 1

2 EFFECT OF AUG-PROXIMAL SEQUENCES ON TRANSLATION IN
   LEISHMANIA .................................................................................................................. 50

3 NON-RANDOM DISTRIBUTION OF TRI-NUCLEOTIDES IN THE AUG-
   PROXIMAL REGION OF LEISHMANIA GENES ...................................................... 111

4 DISCUSSION ................................................................................................................... 148

APPENDIX

A CYSTEINE-LESS GPI-PHOSPHOLIPASE C IS INHIBITED
   COMPETITIVELY BY A SULFHYDRYL REAGENT: GLYCO-MIMICRY
   BY PARA-CHLOROMERCURIPHENYL SULFONATE .............................................. 156
1 Leishmania

1.1 Leishmaniasis

*Leishmania* species, belonging to the Order Kinetoplastida and the Family Trypanosomatidae, are the etiological agents of the disease leishmaniasis. The World Health Organization estimates that over 12 million people are infected with the protozoan parasites, and ten percent of the world's population is at risk of contracting leishmaniasis. The disease is endemic to subtropical and tropical regions, as dictated by the presence of its insect host, the phlebotomine sandfly.

*Leishmania* infections manifest themselves in three forms: visceral, cutaneous, and mucocutaneous leishmaniasis. Invasion of the mononuclear phagocyte system by *L. donovani*, *L. infantum*, or *L. chagasi* results in visceral leishmaniasis or kala-azar [1]. Following an asymptomatic period, the fatal disease is marked by symptoms of fever, cachexia, and hepatosplenomegaly. *L. tropica*, *L. major*, and *L. aethiopica* cause cutaneous leishmaniasis, which is characterized by lesions of the skin. Mucocutaneous leishmaniasis occurs when infection with *L. mexicana* or *L. braziliensis* disseminates from the dermis to the mucosa, often leading to destruction of the nose, mouth, and throat [1].

The most common treatment for leishmaniasis has been pentavalent antimony (sodium stibogluconate and meglumine antimoniate), while
amphotericin B and pentamidine isethionate have also been used [2]. However, severe toxicity and difficult administration are problems for most anti-leishmanial therapies. A new drug with lower toxicity, miltefosine, resulted in a high cure rate for visceral leishmaniasis in clinical trials. Currently, effective vaccines for leishmaniasis are not available.

1.2 *Leishmania* Life Cycle

*Leishmania* has both vertebrate and insect hosts [1]. The flagellated extracellular promastigote lives in phlebotomine sandflies, and the spherical intracellular amastigote is found in vertebrate mononuclear phagocytes (macrophages) (Figure 1.1). The sandfly acquires the parasite while taking a bloodmeal containing amastigotes. Once inside the insect, digestion releases the amastigotes from macrophages. Two cell surface molecules, the glycoconjugate lipophosphoglycan (LPG) and the metalloprotease gp63, help protect the parasites from hydrolytic enzymes in the gut. The amastigotes differentiate into procyclic promastigotes, which replicate by binary fission and attach to the midgut epithelium to avoid excretion. Procyclic promastigotes differentiate into non-dividing metacyclic promastigotes, which detach from the midgut and migrate to the mouthparts. An infected sandfly inoculates the skin of a vertebrate host with metacyclic promastigotes during a bloodmeal (reviewed in [1]).

Complement is deposited onto the parasite in the vertebrate bloodstream, but LPG and gp63 prevent complement-mediated lysis from occurring. Binding and activation of complement to LPG, gp63, and other cell surface molecules
promotes phagocytosis of the metacyclic promastigotes into macrophages. Once inside the macrophage, parasite-containing phagosomes fuse to lysosomes to form a secondary lysosomes (phagolysosomes) within which the metacyclic promastigotes differentiate to amastigotes. Inside the vertebrate cell, LPG and gp63 also protect the parasites from hydrolytic enzymes of the phagolysosome and macrophage oxidative burst. The amastigotes divide until their proliferation ruptures the macrophage, releasing the parasites into the bloodstream to infect more cells (reviewed in [1]).

2 Genome Organization

Although Leishmania are diploid, they replicate by binary fission and do not possess a sexual stage. Their genome is approximately 3.55 megabases (Mb) in size and is 58-60% GC (reviewed in [3]). Leishmania possess 34-36 haploid chromosomes, depending on the species. The chromosome length varies due to differences in the number of repeated sequences (e.g., sub-telomeric repeats, telomeres, or multicopy genes) caused by amplification and deletion events [4]. Leishmania can maintain circular and linear minichromosomes created in response to stress such as drug selection [5-7]. For example, a portion of the of the genome containing the dihydrofolate reductase/thymidylate synthetase (DHFR/TS) gene is amplified to create a 30 kilobase (kb) “R” circle in response to methotrexate selection [5, 7].

The genes on Leishmania major chromosome 1 have an unusual organization [8]. The coding regions are arranged in clusters with relatively short intergenic regions and no introns. The first 29 coding regions are found on one
strand of the DNA, while the next 50 are on the opposite strand [8]. This organization is “divergent” because transcription occurs toward the telomeres (Figure 1.2) [9]. In contrast, the arrangement of genes on chromosome 4 is “convergent” with genes transcribed away from the telomeres [9]. Similar patterns are found in other trypanosomatid genomes. Clustering of *Leishmania* genes on the same DNA strand is reminiscent of prokaryotic operons [4]. However, clustered *Leishmania* genes do not function in the same metabolic pathways, and they are usually not regulated in a similar fashion [4].

3 Transcription

3.1 Polycistronic Transcription

Although most eukaryotes transcribe monocistronic heterogeneous nuclear RNA (hnRNA), trypanosomatids produce polycistronic pre-mRNA. This unique aspect of their gene expression is facilitated by the clustered arrangement of genes [10, 11]. Maturation of individual mRNAs then occurs by two pre-mRNA processing reactions: trans-splicing and polyadenylation [12] (see section 4).

Trypanosomatid polycistronic transcription was discovered in *Trypanosoma brucei* when parasites were subjected to UV irradiation [10]. This treatment resulted in the accumulation of a 60 kilobase (kb) unit containing the gene for variant surface glycoprotein 221 (VSG 221) as well as several other protein-coding genes [10]. Further evidence was provided by detection of unprocessed polycistronic tubulin pre-mRNA in *T. brucei* following heat shock [11]. Both UV irradiation and heat shock result in the accumulation of pre-mRNA
processing intermediates derived from polycistronic transcription units. Polycistronic transcription was later found in other trypanosomatids. In *Leishmania*, polycistronic pre-mRNAs of 100-500 kb have been found [9].

### 3.2 RNA Polymerases

Trypanosomatids utilize three types of RNA polymerases. RNA polymerase II (RNAP II) transcribes mRNAs and small nuclear RNAs, and RNA polymerase III (RNAP III) produces tRNAs and other small RNAs. Interestingly, RNA polymerase I (RNAP I) is used not only for transcribing ribosomal RNA (rRNA), but also for transcribing protein-coding genes (reviewed in [4, 12]). This marks another significant difference between transcription in trypanosomatids and other eukaryotes (which only use RNAP II to transcribe protein-coding genes).

### 3.3 Promoters

The polycistronic nature of transcription in the trypanosomatids has made the identification of promoters challenging. RNA polymerase I (RNAP I) promoters for abundant cell surface proteins in *T. brucei* were the first to be characterized. Transcription of variant surface glycoprotein (VSG) and procyclic-acidic repetitive protein (PARP) genes was shown to be insensitive to the toxin α-amanitin and the detergent Sarkosyl, indicating utilization of RNAP I promoters [13, 14].

Most eukaryotic RNAP I promoters have two important regions: (i) an essential core region including the transcription start site, and (ii) an upstream control element. RNAP I promoters for VSG, PARP, and rRNA in *T. brucei* have
a bipartite core region, consisting of two elements, box 1 and box 2 (Figure 1.3) [15]. The PARP and rRNA promoters also possess upstream elements, but the VSG promoter does not [16]. Despite the organizational similarities, the promoter sequences are not well conserved. RNA polymerase I promoters have also been identified for rRNA in *Leishmania* [17-19].

Three features are generally found in RNAP II promoters in most eukaryotes: (i) a core consisting of an initiator sequence including the transcription start site, (ii) a TATA box with the TATAAAA sequence, and (iii) upstream control elements. RNAP II promoters for protein-coding genes in the trypanosomatids have not been well characterized. The only clearly defined RNAP II promoters are for the spliced leader (SL) RNA gene [20-23]. The SL RNA promoters contain three elements: an initiator and two upstream control elements (UCEs). One of the UCEs contains a conserved GAC trinucleotide, but TATA boxes have not been found [24]. Other putative RNAP II promoters have been reported, but their functionality has not been confirmed [4, 25].

It has been speculated that RNAP II transcription initiation may be non-specific in *Leishmania* because reporter genes can be expressed in the absence of recognizable promoters as long as pre-mRNA processing signals are present [4, 26-30] (see Section 4). However, a region hypothesized to contain a bidirectional promoter for *Leishmania major* chromosome 1 was recently described [25]. Interestingly, transcription of the entire chromosome initiates from a 100 bp sequence within the intergenic region between the 29 genes on one DNA strand and the 50 genes on the other (see Section 1.2, Figure 1.2) [8].
The sequence does not contain a TATA box and does not share sequence similarity with intergenic regions from other *L. major* chromosomes [25].

### 3.4 Transcription Factors

Few putative trypanosomatid transcription factors have been identified experimentally. The *Leptomonas seymouri* SL RNA promoter consists of an initiator and two upstream elements, PBP-1E and PBP-2E [24]. A putative transcription factor, promoter binding protein (PBP-1), associates with the PBP-1E element of the SL RNA promoter [24]. PBP-1 consists of three subunits, one of which appears to be related to a subunit of an small nuclear RNA (snRNA) transcription factor [31]. A 57 kiloDalton (kDa) subunit of PBP-1 was shown to be required for efficient transcription of SL RNA in an *in vitro* system [31].

The proximal sequence element (PSE) of the *T. cruzi* SL RNA promoter is bound by a 45 kDa PSE promoter-binding protein (PPB1) [32]. Overexpression of PPB1 *in vivo* resulted in enhanced transcription of SL-RNA [32]. *In vitro* transcription systems, which are available in some trypanosomatid systems, will aid in the identification and characterization of putative transcription factors [33, 34].

### 4 Pre-mRNA Processing

While most eukaryotic heterogeneous nuclear RNA (hnRNA) is monocistronic, in trypanosomes the corresponding transcripts are polycistronic [10, 11]. Maturation of these transcripts into mRNA occurs by two pre-mRNA processing events, trans-splicing and polyadenylation (Figure 1.4) [12].
Evidence of cis-splicing has been found for only one trypanosomatid gene, poly(A) polymerase [35].

4.1 Trans-splicing

Trans-splicing is an intermolecular reaction whereby a highly conserved 39-41 nucleotide sequence called the spliced leader (SL) is added to the 5' UTR of a pre-mRNA (Figure 1.5A) [36, 37]. SL is a “pseudo exon” at the 5' end of an 80-140 nucleotide precursor SL RNA [38]. SL RNA contains a highly methylated “CAP4” structure at its 5' terminus [39, 40]. There are approximately 150 copies of the SL RNA gene in tandem arrays in the *Leishmania* genome [41]. SL sequences are identical within a species. All trypanosomatid nuclear mRNAs have SL at their 5' end [42].

Trans-splicing is a two step reaction directed by a GU dinucleotide at the 5' splice site of the SL RNA and on the pre-mRNA by (i) a polypyrimidine tract, (ii) a branch site A, and (iii) an AG dinucleotide at the 3' splice site of the 5' UTR (Figure 1.5A) (reviewed in [15, 37]). In the first step of the reaction, the branch point A of the pre-mRNA makes a nucleophilic attack on G at the 5' splice site of the SL RNA. The result is a 2'-5' branched Y intermediate. In the second step, a transesterification reaction ligates the SL “exon” onto the 3’ splice site (splice acceptor site) of the pre-mRNA. The Y intermediate is then debranched and degraded [43].

Trans-splicing was first discovered in the trypanosomatids and later found in other eukaryotes, including *C. elegans, Schistosoma mansoni, Euglena, Drosophila*, and plant chloroplasts [44-48].
Despite its ubiquity, the function of the trypanosomatid spliced leader sequence has not been elucidated. Aside from its role in pre-mRNA maturation, SL may be important for mRNA transport, stability, and translatability. SL provides a cap, but there is no evidence that this is critical for translation in *Leishmania*. In fact, cap-independent translation can occur in the parasite [49], and a cap is not always essential for translation in other eukaryotes [50-55]. It is known that SL is required for expression of reporter proteins in trypanosomatids [4, 26-30], but whether this is indicative of a role in protein synthesis has not been demonstrated experimentally.

4.2 Polyadenylation

In trypanosomatids, polyadenylation is coupled to trans-splicing [56]. Both processing events depend on a common polypyrimidine tract that directs trans-splicing of the downstream gene and facilitates polyadenylation of the upstream gene (Figure 1.5B) [57]. The site of poly(A) tail addition is not determined by a consensus sequence as in other eukaryotes, but instead is influenced by the location of the downstream splice acceptor site AG (see Section 4.1) [56-58]. Polyadenylation occurs 200-700 nucleotides upstream of splice acceptor sites in *Leishmania* [3, 56], and 100-200 nucleotides upstream of splice acceptor sites in *T. brucei* [59]. In trypanosomatids, multiple polyadenylation sites downstream of a coding sequence are possible [56, 59].

Eukaryotic poly(A) tails can affect message stability, in part by association of poly(A)-binding protein (PABP) [60, 61]. Trypanosomatid poly(A) tails are also
thought to be involved in mRNA stability, although there is no experimental evidence for this role. PABP I of *Leishmania* could mediate this function [62].

In other eukaryotes, poly(A) tails are also involved in the regulation of protein synthesis. When ribosomes or translation factors are limiting, a poly(A) tail can act synergistically with a 5’ m7G cap to enhance translation [63, 64]. The two ends of a message interact through association of eukaryotic initiation factor 4G (eIF4G) with both PABP and cap-binding eukaryotic initiation factor 4E (eIF4E) [65]. It is hypothesized that circularization of mRNA [66] could enhance translation by recycling of the 40S subunit (reviewed in [67]). However, a poly(A) tail is not always required for translation [68-74] or for the activity of PABP [75] in other eukaryotes. It is not known whether poly(A) tails affect protein synthesis in trypanosomatids.

5 Translation

5.1 Ribosome Structure

Ribosomes are ribonucleoprotein complexes responsible for translating mRNA. Eukaryotic ribosomes are comprised of two subunits, a small 40S subunit and a large 60S subunit. The 40S subunit serves as a decoding center, mediating the interaction between mRNA and tRNA. The large 60S subunit harbors the peptidyl transferase activity, catalyzed by ribosomal RNA (rRNA) [76, 77]. During translation initiation, the two subunits are brought together to form an 80S ribosome (see Section 5.2). The ribosome has three sites for binding tRNAs: (i) the aminoacyl site (A site) where charged aminoacyl-tRNAs are introduced, (ii) the peptidyl site (P site) for the growing polypeptide chain bound
to tRNA (peptidyl-tRNA), and (iii) the exit site (E site) where uncharged tRNAs leave the ribosome (reviewed in [76, 77]).

The archetypal eukaryotic ribosome has four major rRNAs: an 18S rRNA in the small subunit, and 5S, 5.8S, and 28S rRNAs in the large subunit. In *Leishmania*, the large subunit 28S rRNA is fragmented into two large molecules (a and b) and four small molecules (c, d, e, and x) [78] (Figure 1.6). These molecules are analogous to domains of 28S rRNA [79]. *Leishmania* also have 18S, 5S, and 5.8S rRNA species.

### 5.2 Translation Initiation

The process of eukaryotic translation initiation involves 40S and 60S ribosomal subunits, mRNA initiator tRNA (Met-tRNA$_i$), and eukaryotic initiation factors (eIFs) (reviewed in [80, 81]). The mRNA often has a 5' methyl-7-guanosine cap (5' m7G cap) and a 3' poly(A) tail. Translation begins with formation of a ternary complex between eIF2 (a GTP-binding protein), GTP, and Met-tRNA$_i$ (Figure 1.7). The ternary complex then associates with the 40S subunit, along with eIF3, eIF1, and eIF1A to form the 43S complex [82].

As the 43S complex is formed, other initiation factors assemble on the mRNA to facilitate the recruitment of the small subunit (Figure 1.7). The eIF4F complex (comprised of eIF4E, eIF4A, and eIF4G), eIF4B, and poly(A)-binding protein (PABP) associate with the mRNA. The 5' m7G cap of the mRNA is bound by eIF4E, and the RNA helicase activity of eIF4A (stimulated by eIF4B) helps unwind secondary structures in the 5' untranslated region (5' UTR). Acting as a scaffold, eIF4G binds to several initiation factors including eIF4E at the 5'
cap and PABP at the 3’ poly(A) tail, effectively circularizing the mRNA [65, 66]. The association of eIF4G with eIF3 may help bring the 43S complex to the 5’ end of the mRNA [83]. The 43S subunit can also bind directly to internal ribosome entry sites in the 5’ UTR (see Section 5.4).

Once in contact with the mRNA, the 43S complex is hypothesized to scan the mRNA linearly until it reaches the translation initiation codon (AUG), forming the 48S complex (Figure 1.7) (reviewed in [80, 81]). Movement of the 43S to the AUG is mediated by eIF1 and eIF1A [82, 84]. Once the small subunit is positioned at the translation initiation codon, eIF5 triggers the hydrolysis of GTP bound to eIF2, leading to the release of eIF2 and eIF3 [85]. Then eIF5B hydrolyzes its bound GTP to facilitate the joining of the 60S subunit to the 48S complex [85, 86]. This results in the formation of the 80S ribosome with the Met-tRNA\textsubscript{i} positioned at the peptidyl-tRNA site (P site) of the small subunit.

5.3 The Scanning Model of Translation Initiation

The Scanning Model is one mechanism describing ribosome selection of a translation initiation start site. The model proposes that the 43S initiation complex associates with the m7G cap at the 5’ end of mRNA. It scans the message in the 5’ to 3’ direction until it encounters the first start codon in a “favorable” context (reviewed in [80]).

The Kozak box is a consensus sequence that was originally identified in an alignment of the -12 to -1 regions of 699 vertebrate mRNAs [87]. It consists of seven nucleotides that abut the start codon (GCCA/GCCaug) [87]. From mutational analysis of this sequence, it was determined that A at -3 and G at +4
are most important for start site selection [88, 89]. The context of an AUG codon has been classified as “weak” or “strong” based on the nucleotides occupying these positions [80]. In contrast to the Kozak box, an alignment of 5’ UTRs from 96 *Saccharomyces cerevisiae* genes resulted in an A-rich consensus sequence AAAAAAaugUCU [90].

The significance of the Kozak box was reevaluated in a study of 2195 human mRNAs [91]. Only well-characterized transcripts were used, rather than coding regions predicted from genomic sequence alone. Only 46% of cytosolic mRNA sequences and 42% of secretory mRNA sequences examined have an A or a G at -3 and a G at +4, indicating that over half the transcripts do not utilize the “favorable” context defined by the Kozak box [91]. In fact, 76% of the transcripts have three or more mismatches from a CCACCaugG sequence [91]. Upstream AUGs were found in 65% of the transcripts, and 47% of these upstream AUGs were in a similar or more favorable context as compared to the authentic AUG [91]. This suggests that use of the first AUG in a favorable context rule may not always hold true.

### 5.4 Other Mechanisms of Translation Initiation

Besides the Scanning Model, translation initiation can occur by three additional mechanisms: (i) “leaky” scanning, (ii) reinitiation, and (iii) internal initiation [80]. Leaky scanning occurs when the 40S small ribosomal subunit bypasses the first AUG in the 5’ UTR of an mRNA and instead initiates translation at a downstream AUG. In reinitiation, after an upstream open reading frame (uORF) is translated (i) the 40S subunit remains associated with the
mRNA, (ii) continues scanning, (iii) reacquires the ternary complex, and (iv) initiates translation at a downstream reading frame (see Section 7.2). Internal initiation occurs by cap-independent ribosome recruitment to an internal ribosome entry site (IRES) in the 5' UTR (reviewed in [92]) (see Section 7.2).

### 5.5 Translation Elongation & Termination

Protein synthesis proceeds with elongation (reviewed in [77, 93]), which requires tRNA molecules charged with cognate amino acids (aminoacyl-tRNAs or aa-tRNAs). A tRNA molecule has an L shape with an anticodon (consisting of three bases complementary to an mRNA codon) at the end of one loop. During charging of an amino acid to a tRNA, one of 20 specific aa-tRNA synthetases catalyzes the formation of an ester bond between the amino acid encoded by the anticodon and the 3' end of the tRNA molecule.

In the first step of elongation, a ternary complex of elongation factor 1A (eEF1A, a GTP binding protein), GTP, and an aa-tRNA is brought to the aminoacyl site (A site). If the decoding center of the 40S subunit detects a match between the tRNA anticodon and the codon of the message, eEF1A hydrolyzes its GTP and is released, leaving the aa-tRNA in the A site. The peptidyl transferase activity of the large subunit 28S rRNA catalyzes peptide bond formation between the growing polypeptide at the P site and the amino acid at the A sites. The reaction results in the transfer of the peptidyl-tRNA from the P site to the A site, and leaves an uncharged tRNA at the P site. The ribosome advances to the next codon of the mRNA in a process called “translocation”, which is facilitated by the hydrolysis of GTP bound to eEF2. During
translocation, the peptidyl-tRNA is transferred from the A site to the P site, and the discharged tRNA at the P site is moved to the E site.

The steps in elongation are repeated until a termination codon is positioned at the A site (reviewed in [94]). Termination codons (UAG, UAA, or UGA) are recognized by eukaryotic release factors, which facilitate the cleavage and release of the polypeptide at the P site and discharge the tRNA at the exit (E) site.

6 Control of Trypanosomatid Gene Expression

In trypanosomatids, there is little or no control at the level of transcription initiation, because most promoters are constitutively active. As a result, post-transcriptional mechanisms for controlling gene expression dominate. It is well documented that protein abundance frequently does not correlate with mRNA levels in these organisms [95-98]. For example, in Trypanosoma cruzi, steady state levels of cruzipain (a cysteine protease) mRNA are the same during the life cycle, but the amount of protein is much higher in the insect stage [96]. Pre-mRNA processing, mRNA localization, transcript stability, and mRNA translatability may contribute to regulation after transcription.

Post-transcriptional control mechanisms allow mature transcripts from the same polycistronic pre-mRNA to be translated at different levels [99-102]. For example, the variant surface glycoprotein (VSG) mRNA from T. brucei is over a 100-fold more abundant than mRNAs for expression site-associated genes (ESAGs) originating from the same polycistronic transcript [101]. Additionally, mature transcripts from the same polycistronic pre-mRNA can give rise to varying
amounts of protein [103, 104]. The stability and translation of mature transcripts from a polycistronic pre-mRNA are not co-regulated, and individual control is likely to involve untranslated regions.

7 Untranslated Regions of mRNA

Untranslated regions (UTRs) are found upstream and downstream of the protein coding sequence of an mRNA. Typically, the 5’ UTR extends from the m7G cap to the translation initiator, while the 3’ UTR begins at the translation termination codon and ends at start of the poly(A) tail. Sequences in the untranslated regions can affect message stability, mRNA translatability, transcript localization, and polyadenylation of mRNA (reviewed in [105]) (Figure 1.8).

7.1 Untranslated Regions in Trypanosomatids

In trypanosomatids, post-transcriptional mechanisms are crucial for controlling gene expression. UTRs are likely to be important for this type of regulation. Studies of trypanosomatid UTRs have focused on their role in differentiation from one cell type to another (e.g., Leishmania insect stage promastigote to vertebrate stage amastigote). Little is known about the contribution of UTRs to gene expression in a single stage.

Untranslated regions can affect mRNA stability during differentiation. Expression of amastigote-specific A2 genes in Leishmania donovani is controlled by the 3’ UTR, which facilitates mRNA stability in amastigotes, but leads to degradation in promastigotes [106, 107]. In T. brucei, the 3’ UTRs of variant surface glycoprotein (VSG) and procyclic acidic repetitive protein (PARP)
contribute to stage-specific expression by increasing transcript stability in the vertebrate stage and insect stage, respectively (reviewed in [4, 12]).

UTRs that influence other levels of control besides mRNA stability have been identified. For example, a 450 nucleotide sequence in the 3’ UTR of *L. major* amastin surface proteins confers preferential amastigote stage expression without affecting mRNA stability, perhaps by affecting translation [108]. Interestingly, similar sequences are also found in several other *Leishmania* mRNAs [108].

While it has been demonstrated that 3’ UTRs play a role in stage-specific expression of several trypanosomatid genes, 5’ UTRs in general have received little attention. (Most studies of trypanosomatid 5’ UTRs have focused on the identification of sequences important for trans-splicing (see Section 4.1)). In *Crithidia fasciculata*, octamer consensus sequences in the 5’ UTR of two transcripts increase their stability just before S phase [109, 110]. The effect may be mediated by the binding of specific protein factors to the element [111-113].

**7.2 Effects of 5’ UTRs on Translation**

Cis-acting elements in a 5’ UTR can regulate translation. An iron response element (IRE) in the 5’ UTR of mRNAs involved in iron metabolism forms a hairpin with a bulged cytidine (reviewed in [114]). When iron levels are low, iron regulatory proteins (IRP-1 and IRP2) bind IRE and inhibit translation by blocking the recruitment of the 43S initiation complex [115]. In response to high levels of iron, IRP dissociates from IRE, allowing protein synthesis to occur. The 5’ terminal oligopyrimidine tract (TOP) is another example of a cis-element found
in 5' UTRs. Many mammalian mRNAs encoding proteins for the translational apparatus contain TOPs (reviewed in [116]). TOP elements consist of a C at the 5' m7G cap site followed by 4-14 pyrimidines. In response to events such as growth arrest, this sequence facilitates coordinated repression of translation so that energy can be conserved. Although potential binding proteins have been suggested, the trans-acting factors responsible for inhibiting translation of TOP mRNAs have not been identified [116].

Internal ribosome entry sites (IRESs) in the 5' UTR allow cap-independent translation initiation (see Section 5.4). IRESs were first discovered in picornaviruses [117, 118] and later found in other viruses as well as cellular mRNAs such as immunoglobulin heavy chain-binding protein [119]. These sequences in the 5' UTR of an mRNA form a scaffold for direct or factor-mediated binding of the 40S subunit upstream of the AUG initiator, allowing internal initiation of translation.

IRESs are defined by their secondary and tertiary structure rather than by nucleotide sequence. Most IRESs form a common Y-shaped stem-loop structure upstream of the initiation codon [120]. Trans-acting factors that facilitate IRES-dependent translation have been identified (reviewed in [82]). For example, pyrimidine tract-binding protein (PTB) is thought to be important for the function of picornavirus IRESs [121-124]. It is hypothesized that binding of PTB to an IRES helps stabilize the sequence for 43S complex binding [82].
Internal initiation of translation can also occur in *Leishmania*. Leishmania RNA virus 1 possesses an IRES in its 5' UTR, which is used for cap-independent translation initiation in the parasite [49].

Upstream ORFs (uORFs) in a 5' UTR can play a role in translation regulation by mechanisms involving reinitiation (see Section 5.4). After translating a uORF, a 40S subunit may resume scanning and reinitiate at a downstream AUG or leave the mRNA without translating additional ORFs. For example, *Saccharomyces cerevisiae* GCN4 mRNA encodes a transcription factor for genes involved in amino acid biosynthesis and tRNA charging (reviewed in [125]). The 5' UTR of GCN4 mRNA contains four uORFs that inhibit its translation. Under normal conditions, ribosomes dissociate from the mRNA after terminating translation of uORF4, thus preventing GCN4 protein synthesis. During amino acid starvation, the 40S subunit is not reloaded with the ternary complex in time to translate uORF4. Reinitiation occurs at the downstream GCN4 ORF instead, allowing translation of GCN4p.

### 7.3 Effects of 3' UTRs on Translation

3' UTRs can play a role in regulating translation initiation. For example, mammalian 15-lipoxygenase (15-LOX), an enzyme involved in red blood cell maturation, has a differentiation control element (DICE) in its 3' UTR that facilitates tissue-specific translation of the enzyme [126]. DICE is bound by heterogeneous nuclear ribonucleoprotein (hnRNP) K and hnRNP E1 [127], which prevent the formation of an 80S ribosome on the 15-LOX mRNA in non-erythroid tissues [128].
Translation can also be regulated after the initiation step by sequences in 3’ UTRs. In *C. elegans*, the *lin-4* gene encodes a small noncoding RNA that binds to complementary sequences in the 3’ UTR of *lin-14* [129]. This prevents translation of *lin-14* after initiation [130]. Nanos (nos) mRNA from *Drosophila* has a translation control element (TCE) in its 3’ UTR. When bound by Smaug protein, translation of nos mRNA is inhibited, although initiation of translation is not [131]. In both cases, ribosomes associated with *lin-14* and nos mRNAs were shown to be translationally active because an aa-tRNA analog, puromycin, was able to displace the transcripts from polysomes.

Cytoplasmic polyadenylation helps activate maternal mRNAs during vertebrate development. A cytoplasmic polyadenylation element (CPE) in the 3’ UTR of mRNAs such as *Xenopus* c-mos is important for this process [132, 133]. These mRNAs are translationally repressed when cytoplasmic polyadenylation element binding protein (CPEB) binds CPE and Maskin protein [134, 135]. Maskin competes with eIF4E for binding to eIF4G, thereby inhibiting initiation of protein synthesis [135]. Phosphorylation of CPEB causes it to recruit the polyadenylation complex and poly(A) tail elongation ensues [136]. The exact mechanism by which increasing poly(A) tail length stimulates translation is not known, but it is thought to involve the binding of additional PABP [137].

### 7.4 Resources for UTR analysis

UTRdb, UTRsite and UTRscan are databases available for analysis of untranslated regions (http://bighost.area.ba.cnr.it/BIG/UTRHome/) [105]. UTRdb has a non-redundant collection of 5’ and 3’ UTR sequences from eukaryotic
mRNAs. The UTRs were obtained from EMBL/Genbank and include human, rodent, invertebrate, plant, fungal, and viral sequences. UTRsite is a database of experimentally defined cis elements of 5’ and 3’ UTRs. Each entry provides a description of the pattern’s function as well as its primary and secondary structure. UTRscan allows a user to submit and search input sequences for the presence of cis-elements in UTRsite.
REFERENCES


[71] Adesnik, M., Salditt, M., Thomas, W. & Darnell, J. E. (1972) Evidence that all messenger RNA molecules (except histone messenger RNA) contain Poly (A) sequences and that the Poly(A) has a nuclear function, *J Mol Biol.* 71, 21-30.


Identification and overexpression of the A2 amastigote-specific protein in

expression in Leishmania mediated by a conserved 3'-untranslated region

Sequences within the 5' untranslated region regulate the levels of a
16, 6724-6735.

levels in the trypanosomatid *Crithidia fasciculata*, *Nucleic Acids Res.* 25,
3281-3289.

fasciculata* contain a factor(s) that binds to the 5'-untranslated regions of
TOP2 and RPA1 mRNAs containing sequences required for their cell cycle

trans elements involved in the cell cycle regulation of multiple genes in
*Crithidia fasciculata*, *Mol Cell Biol.* 19, 6174-82.

of the *Crithidia fasciculata* mRNA cycling sequence binding proteins, *Mol
Cell Biol.* 21, 4453-9.

metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide,
and oxidative stress, *Proc Natl Acad Sci U S A.* 93, 8175-82.

binding of the 43S translation pre-initiation complex to ferritin and eALAS


C. & Wimmer, E. (1988) A segment of the 5' nontranslated region of
encephalomyocarditis virus RNA directs internal entry of ribosomes during


regulatory complex controls 60S ribosomal subunit joining, *Cell.* 104, 281-90.


**Figure 1.1: *Leishmania* life cycle** (from [138]). During a bloodmeal, an infected sandfly inoculates the vertebrate host with metacyclic promastigotes. In the bloodstream, the parasites are phagocytosed by macrophages. The metacyclic promastigotes differentiate to amastigotes and divide. Their proliferation ruptures the macrophage, releasing parasites into the bloodstream to infect more cells. The sandfly acquires amastigotes while taking a bloodmeal on an infected vertebrate host. Inside the insect, amastigotes differentiate into procyclic promastigotes, divide, and attach to the midgut epithelium. Procyclic promastigotes differentiate into metacyclic promastigotes, which detach from the midgut and migrate to the sandfly mouthparts.
Figure 1.2: Gene organization in *Leishmania major* (from [9]). Chromosome 1 (Chr 1) has a “divergent” arrangement of genes. The two clusters of genes (thick lines) are transcribed toward telomeres. On chromosomes 3 and 4 (Chr 3 and Chr 4) gene clusters are transcribed away from the telomeres in a “convergent” manner.
Figure 1.2
Figure 1.3: Structure of *T. brucei* RNA polymerase I promoters (from [15]).

Trypanosomatids can use RNA polymerase I to transcribe protein coding genes. The promoters have a two domain structure: (i) a bipartite core consisting of box 1 and box 2, and (ii) upstream control elements. RIB = ribosomal RNA promoter, PRO = procyclic acidic repetitive protein (PARP) promoter, VSGB = variant surface glycoprotein promoter.
Figure 1.3
Figure 1.4: Outline of trypanosomatid gene expression. Promoters that have been analyzed are always on, and the resulting transcripts are polycistronic. Pre-mRNAs mature into monocistronic transcripts by trans-splicing and polyadenylation.
Figure 1.4

Most promoters are always active.

DNA

Transcription produces polycistronic pre-mRNA.

Polycistronic transcription unit

Trans-splicing and polyadenylation define 5’ and 3’ ends.

Monocistronic mature mRNAs
Figure 1.5: Maturation of trypanosomatid pre-mRNAs by trans-splicing and polyadenylation (from [15]).  

**A)** Trans-splicing is an intermolecular reaction in which a conserved sequence called the spliced leader (SL) is added to the 5’ UTR of a pre-mRNA.  5’ SS = 5’ splice site, 3’ SS = 3’ splice site (splice acceptor site), BP = branch point, DebrEnz = debranching enzyme.  

**B)** Polyadenylation and trans-splicing are coupled; both events are directed by a single polypyrimidine tract between adjacent genes.
Figure 1.5

A)

5' SS
SL RNA

● SL GU

+ A−AG

pre-mRNA

BP 3' SS

GU

A−AG

DobrEnz

GU

+ A−AG

B)

Polyadenylation and Trans-splicing

pre-mRNA

mature mRNA 5' end

mature mRNA 3' end

AG Intergenic sequences
Figure 1.6: Comparison of rRNA from trypanosomatids and other eukaryotes. The typical eukaryotic ribosome has four major rRNA species: an 18S rRNA in the 40S small subunit, and 5S, 5.8S, and 28S rRNAs in the 60S large subunit. In *Leishmania*, the large subunit 28S rRNA is fragmented into two large molecules (a and b) and four small molecules (c, d, e, and x).
Figure 1.6

**Eukaryotic ribosomal RNA**

- 60S
- 40S
- 5S rRNA, 5.8S rRNA
- 28S rRNA
- 18S rRNA

**Trypanosomatid ribosomal RNA**

- 60S
- 40S
- 5S rRNA, 5.8S rRNA
- a, b, c, d, and x rRNA
- 18S rRNA
Figure 1.7: Eukaryotic translation initiation (from [81]). A ternary complex of eIF2, GTP, and Met-tRNA\textsubscript{i} associates with the 40S subunit, eIF3, eIF1, and eIF1A to form the 43S complex. On the mRNA, eIF4E binds the cap, eIF4A (assisted by eIF4B) unwinds the secondary structure, and eIF4G associates with PABP at the poly(A) tail, eIF4E, and other initiation factors. The 43S complex is recruited to the 5’ end of the mRNA and may scan in the 5’ to 3’ direction until it reaches the translation initiation codon (AUG), where it forms the 48S complex. Hydrolysis of GTP bound to eIF2 is triggered by eIF5, releasing eIF2-GDP and eIF3. The 60S subunit joins the 48S complex to form the 80S ribosome. eIF = eukaryotic initiation factor, PABP = poly(A) tail binding protein.
Figure 1.7
Figure 1.8: Effects of eukaryotic untranslated regions on gene expression (from [105]). 5’ and 3’ UTRs can influence translation, localization, stability, and polyadenylation of mRNA. Some of these mechanisms are mediated by UTR secondary structure and/or by association of the UTR with proteins or RNA. m7G = 5' methyl-7-guanosine cap, uORF = upstream open reading frame, IRES = internal ribosome entry site, CPE = cytoplasmic polyadenylation element.
Figure 1.8
CHAPTER 2

EFFECT OF AUG-PROXIMAL SEQUENCES ON TRANSLATION IN

LEISHMANIA\textsuperscript{1}

\textsuperscript{1} Stanton, Julie D. and Kojo Mensa-Wilmot. To be submitted to Journal of Biological Chemistry.
ABSTRACT

Gene expression in trypanosomatids is controlled predominantly at the post-transcriptional level. Translation is likely to be a significant point of regulation in *Leishmania*. We explored the contribution of 5' untranslated regions (5' UTRs) to protein synthesis in *Leishmania*, by testing the effect of a short sequences just upstream of the translation initiation codon. The 12-nucleotide sequences contained (i) a polypurine box (AGGAGG), (ii) a spacer (CAG), and (iii) nucleotides in the AUG-proximal region (APR) at positions -3 to -1. Eight 12-mers with different APRs were linked to two reporter genes and studied in stable transfectants of *Leishmania tropica*. Sequences in the -3 to -1 resulted in dramatic differences in protein expression. A 700-fold difference in protein expression was observed with a phospholipase C as a reporter. With a luciferase reporter a 6000-fold difference in protein expression could be detected. Three APRs resulted in robust expression of both reporter proteins: ACC, CCC, and GCC. Variation in the steady-state level of reporter mRNA could not explain the effects of the APRs on protein expression, indicating post-transcriptional action of the sequences. Ribosome recruitment was tested as a possible contributor to the effect of the APRs on protein expression. Most APRs that gave high protein levels also resulted in efficient polysome formation on reporter mRNA. One APR inhibited reporter protein expression although the reporter mRNA associated with polysomes. We conclude that the AUG-proximal region of 5' UTRs can regulate gene expression (positively or negatively) by influencing protein synthesis in *Leishmania*. Interestingly, while probing the nature of
reporter mRNA association with ribosomes, we found that *Leishmania* ribosomes are stable in unusually high concentrations of EDTA (50 mM) and KCl (800 mM). *

*The abbreviations used are: A, adenine; APR, AUG-proximal region; AUG, translation initiation codon; bp, base pairs; C, cytosine; DHFR/TS, dihydrofolate reductase thymidylate synthetase; DNase, deoxyribonuclease; DTT, dithiothreitol, EDTA, (ethylenedinitrilo)tetraacetic acid; E-64, trans-Epoxysuccinyl-L-leucylamido(4-guanidino)butane; G, guanine, G418, geneticin; GPI-PLC, glycosylphosphatidylinositol-specific phospholipase C; HEPES, N-(2-Hydroxyethyl)piperazine-N’-(2-ethanesulfonic acid); LUC, luciferase; ml, micro liter; ml, milliliter; NEO, neomycin phosphotransferase; NP-40, nonidet P-40; p, protein; PBS, phosphate buffered saline; PMSF, phenylmethanesulfonyl fluoride; poly(A+), polyadenylated; poly(A-), non-polyadenylated; poly(A) tail, polyadenylate tail; PCR, polymerase chain reaction; PIC, protease inhibitor cocktail; RNase, ribonuclease; RNasin, ribonuclease inhibitor; rpS8, ribosomal protein S8; RT-PCR, reverse transcriptase polymerase chain reaction; SL, spliced leader; T, thymine; TE, Tris-EDTA; TLCK, N-P-Tosyl-L-lysine chloromethyl ketone; UTR, untranslated region.
INTRODUCTION

Control of gene expression in trypanosomatids is primarily post-transcriptional because most promoters are active by default [1, 2]. Consequently, pre-mRNA processing, transcript stability, and mRNA translatability contribute immensely to gene regulation. Post-transcriptional control of gene expression may be influenced by 5′ or 3′ untranslated regions (UTRs). In kinetoplastid protozoans, the role of 5′ UTRs in mature mRNAs to protein synthesis is unknown.

Heterogeneous nuclear RNA (hnRNA) is polycistronic in trypanosomatids [1, 2]. Conversion of these transcripts into mature mRNA involves trans-splicing and polyadenylation, which generate monocistronic mRNAs [3].

Protein abundance frequently does not parallel mRNA levels in trypanosomatids [4-7]. Genes encoded on the same polycistronic pre-mRNA can give rise to different amounts of protein [8-11]. To logically investigate how mRNAs are used at different efficiencies for protein synthesis in the trypanosomatids, one must decipher regions of mRNA that modulate translation, and explore mechanisms used to control protein synthesis (positively or negatively).

We are interested in determining the effects of 5′ UTRs on translation during a single developmental stage of *Leishmania*. To that end, we tested eight 12-nucleotide sequences [12] just upstream of the translation initiation codon in the context of a 250-nucleotide *Leishmania* dihydrofolate reductase thymidylate synthetase (DHFR/TS) intergenic region [13] (Figure 2.1A). The 12-mers
consisted of (i) a polypurine box (AGGAGG), (ii) a spacer (CAG), and (iii) nucleotides in the AUG-proximal region (APR) at positions –3, -2 and -1 [12]. APRs resulted in up to a 6000-fold difference in reporter protein expression. Steady-state levels of reporter transcripts did not correlate with the amount of reporter protein detected. In most cases, APR stimulation of reporter protein synthesis correlated with efficient ribosome recruitment to mRNA. In one instance, a poorly translated reporter mRNA was associated with polysomes. *Leishmania* polysomes proved to be unusually stable to reagents that disrupt ribosome subunit association in other organisms. These observations represent the first report that AUG-proximal regions of a trypanosomatid 5’ UTR modulate translation.

**MATERIAL AND METHODS**

**Materials**

Restriction enzymes were obtained from New England Biolabs (Beverly, MA), and Amplitaq DNA polymerase was from Perkin Elmer (Boston, MA). Luciferin, T4 DNA ligase, RNase inhibitor, and RQ1 DNase were purchased from Promega (Madison, WI). Fetal bovine serum was obtained from Hyclone (Logan, UT), Medium 199 from Gibco BRL (Grand Island, NY), antibiotic/antimycotic from Mediatech (Herndon, VA), and RPMI media without methionine or cysteine from ICN (Costa Mesa, CA). Protease inhibitors and Titan One Tube RT-PCR reagents were purchased from Roche (Indianapolis, IN). The bicinchoninic acid assay kit and luciferin were obtained from Pierce (Rockford, IL), and cycloheximide and NP-40 detergent from Calbiochem (San Diego, CA).
BrightStar psoralen-biotin, Northern Max, BrightStar Plus membrane, UltraHyb Buffer, Biodetect, Micro Poly(A) Plus, and RNase-free TE buffer were purchased from Ambion (Austin, TX). Biomax MR film was obtained from Kodak (Rochester, NY), and Qiagen Plasmid Midi Kit from Qiagen (Valencia, CA). G418, isopropanol, phenol, chloroform, and phenol/chloroform/isoamyl alcohol were purchased from Fisher (Pittsburgh, PA). All other reagents were obtained from Sigma (St. Louis, MO).

**Plasmid construction**

Glycosylphosphatidylinositol-specific phospholipase C (GPI-PLC) plasmids were constructed by isolating Bam HI fragments containing both APRs (Table 2.1) and GPI-PLC genes from pBluescript II constructs [12]. DNA fragments were ligated into Bam HI digested pX63NEO [13, 14] (Figure 2.1B).

To construct luciferase plasmids, the luciferase (LUC) gene was PCR amplified [15] from plasmid p220S (a gift from Dr. Mary Wilson, University of Iowa) [16] using forward primers that introduced different 12-mers (Table 2.2). Products were digested with Bam HI and ligated into the Bam HI site of pX63NEO [13, 14] (Figure 2.1C).

**Culture of Leishmania**

*Leishmania tropica* strain HOM/IQ/73/LCR-L32 [17] was cultured in Medium 199 (M199) (supplemented with 10% fetal bovine serum, 100 μM adenine, 40 mM HEPES, pH 7.5, 0.0005% hemin, 0.0002% biotin, and 50 U/ml antibiotic antimycotic) [18] at 25°C.
DNA transfection

*Leishmania* growing at a density of $8.0 \times 10^6$ to $1.0 \times 10^7$ cells/ml were washed in phosphate buffered saline (PBS) (136 mM NaCl, 27 mM KCl, 5.3 mM Na$_2$HPO$_4$, 1.7 mM KH$_2$PO$_4$, pH 7.4) and resuspended in electroporation buffer (21 mM Hepes, pH 7.5, 137 mM NaCl, 5 mM KCl, 0.7 mM Na$_2$PO$_4$, 6 mM glucose) to a final concentration of $1.0 \times 10^8$ cells/ml. Four hundred microliters of the cell suspension were added to 10-50 µg of plasmid DNA purified by QIAGEN Plasmid Midi Kit, and transferred to a 2 mm gap disposable electroporation cuvette (Genetronics). The cells were incubated on ice for 10 min and then electroporated at 475 V, 800 µF, and 13 W (one pulse) (BTX Electro Cell Manipulator 600). Parasites were incubated on ice for 10 min, and transferred to a flask containing 5 ml of M199. The cells were incubated for 8-12 hours at 25°C before adding G418 (30 µg/ml) for selection of transfected *Leishmania*. Stable transfectants were obtained after 14-28 days. The cells were not cloned.

GPI-PLC enzyme assay

Stably transfected *Leishmania* growing in 100 µg/ml G418 ($10^8$ cells) were washed in PBS and then lysed in 1 ml hypotonic lysis buffer (10 mM Tris-HCl, pH 8.0, 2 mM EDTA) supplemented with a protease inhibitor cocktail (PIC) of 0.4 U aprotinin, 2.1 µM leupeptin, 0.1 mM TLCK [19, 20]. The cells were incubated on ice for 20 min and then centrifuged for 20 min (14,000 x g, 4°C). The pellet was dissolved in 500 µl of GPI-PLC assay buffer (50 mM Tris-HCl, pH 8.0, 5 mM EDTA, 1.0% NP-40) and incubated on ice for 20 min [20]. After centrifugation for 20 min (14,000 x g, 4°C), 5 µl of the supernatant were used in a GPI-PLC assay
with two µg (in 10 µl GPI-PLC assay buffer) of [³H]myristate-labeled membrane
form variant surface glycoprotein ([³H]mfVSG) [20]. Fifteen µl of GPI-PLC assay
buffer were added for a total volume of 30 µl. The reaction was incubated at
37°C for 20 min, and terminated by addition of 500 µl water-saturated n-butanol.
Four hundred microliters of the upper organic phase were retrieved for liquid
scintillation counting [20]. Total protein content of the lysates was determined
with a bicinchoninic acid assay (Pierce).

**Luciferase enzyme assay**

Stably transfected parasites (1 x 10⁷ cells) growing at a density of 8.0 x
10⁶ cells/ml were washed with 1 ml of PBS, pH 7.4 and stored at -20°C. Frozen
cells were lysed in 200 µl of lysis buffer (100 mM potassium phosphate, pH 7.8,
0.2% Triton X-100, and 2 µg/ml leupeptin) [16, 21]. The lysate was centrifuged
for 10 min (8,000 x g, 4°C), and 5 µl were added to 75 µl of luciferase assay
buffer (25 mM glycylglycine, pH 7.8, 5 mM ATP, pH 7.5, 15 mM MgSO₄) in a 96
well plate [16, 21]. The reaction was initiated by injection of 50 µl of 0.5 mM
luciferin substrate in a luciferase assay buffer (Pierce). Peak light emission was
measured for 1 sec after a 2 sec delay and recorded by a FLUOstar OPTIMA
luminometer (BMG Labtechnologies).

**RNA isolation**

Parasites (1 x 10⁸ cells) harvested at a density of 8.0 x 10⁶ cells/ml were
washed with 10 ml of PBS before lysing with 1 ml of TriReagent (Sigma). Two
hundred microliters of chloroform were added, and the sample centrifuged for 15
min (12,000 x g, 4°C). The upper aqueous phase was retained, and RNA was
precipitated by addition of 0.5 ml isopropanol. The sample was centrifuged for 15 min (14,000 x g, 4°C). The RNA pellet was washed with 75% ethanol and resuspended in RNase-free TE buffer. Typically, 200-400 μg of total RNA was obtained from 1 x 10^8 cells

**Synthesis of probes for RNA analysis**

DNA probes for RNA analysis were made by PCR [15] (Table 2.3). Pairs of probes were specific for *GPI-PLC* [22, 23], S8 ribosomal protein (*rpS8*) [24], *LUC* [25], *gp63* [26], large subunit ribosomal RNA (gamma subunit) [27], or small subunit 18S ribosomal RNA [27] (Table 2.3). PCR products were separated from templates by agarose gel electrophoresis, and isolated from the agarose with homemade spin columns. Columns were made by (i) piercing a hole in the bottom of a 0.8 ml Eppendorf tube, (ii) stuffing the 0.8 ml tube with glass wool, and (iii) placing the 0.8 ml tube in a 1.5 ml Eppendorf tube. Gel slices containing PCR products were placed in the 0.8 ml tube and centrifuged twice for 15 min (8,000 x g, 25°C). DNA was extracted from the eluate with phenol/chloroform/isoamyl alcohol and ethanol precipitation [14].

One μg of DNA in 10 μl of RNase-free TE buffer was labeled with 0.25 μg psoralen-biotin (BrightStar psoralen-biotin, Ambion) by UV irradiation (365 nm) for 45 min. Non-crosslinked psoralen-biotin was twice extracted from the TE buffer with 200 μl water saturated n-butanol. Labeled probes were stored at -80°C. Prior to hybridization, probes in 25 μl RNase-free TE buffer were incubated at 100°C for 10 min and quick-chilled in an ice-water bath.
Northern blotting of GPI-PLC and LUC transcripts

For each clone, 20 μg of total RNA in 10 μl RNase-free TE buffer was denatured for 15 min at 65°C and quick-chilled in an ice-water bath. RNA was separated on a formaldehyde 1% agarose gel (Northern Max Formaldehyde Gel solutions, Ambion), which was stained with ethidium bromide. RNA was vacuum-blotted to BrightStar Plus membrane (Ambion) using a Posiblot Transfer Apparatus (Stratagene) and then crosslinked with UV for one minute at 1.2 kJ in a Stratagene UV crosslinker (Stratalinker 1800). After 2 hours prehybridization in 10 ml of UltraHyb Buffer (Ambion) at 42°C, the membrane was hybridized with approximately 250-300 ng of psoralen-labeled DNA probe for 16-18 hours at 42°C. Probe was detected using the BioDetect system (Ambion), and the membrane exposed to BioMax MR film (Kodak). GPI-PLC, LUC, and rps8 mRNAs were quantitated directly from the membrane using a Genegnome Chemiluminescence Imager and Genetools software (Syngene).

RT-PCR

One μg of Leishmania total RNA was used for Titan™ One Tube RT-PCR (Roche). Synthesis of cDNA was directed by a KCR2, a reverse primer specific to the GPI-PLC gene (Table 2.3). KCR2 was used with LmSL2, a forward primer specific to the Leishmania tropica spliced leader (SL) (Table 2.3) in the PCR step of the reaction. Products were separated on a 1.2% agarose gel and then stained with ethidium bromide [14].
RNA dot blot analysis

Polyadenylated and non-polyadenylated RNA were separated by two rounds of adsorption to Poly(A) Plus columns (Micro Poly(A) Plus, Ambion). After denaturing the RNA at 65°C for 15 min, the samples were quick-chilled in an ice-water bath. One microgram of poly(A+) RNA and 1 μg of poly(A-) RNA were spotted onto a BrightStar Plus membrane (Ambion). Crosslinking, prehybridization, hybridization, and detection were performed as described for Northern blotting, except 60 to 75 ng of probe were used.

Polysome analysis

*Leishmania* (5 x 10⁸ cells) harvested at a density of 8.0 x 10⁶ cells/ml were washed with 5 ml M199 and then with 10 ml PBS (both containing 100 μg/ml cycloheximide). Parasites were resuspended in 750 μl Buffer A (10 mM Tris-HCl, pH 7.4, 300 mM KCl, 10 mM MgCl₂, 1 mM DTT, 100 μg/ml cycloheximide, 1 mM PMSF, 8.5 μg/ml aprotinin, 50 μg/ml leupeptin, 1 μM pepstatin, 50 μg/ml TLCK, 10 μM E-64) supplemented with 1 μl RNasin (40 U) (Promega) [28]. The cells were incubated on ice for 3 min before addition of 125 μl of cell lysis buffer (Buffer A with 1.2% Triton X-100, 0.2M sucrose, 60 U/ml RNasin) [28]. Parasites were further fragmented with a 7 ml Dounce homogenizer (fifteen strokes) and transferred to a 1.5 ml tube and centrifuged for 2 min (14,000 x g, 4°C). The supernatant was added to a 1.5 ml tube containing 100 μl heparin (10 mg/ml) and 1 μl RNasin (40 U), and layered onto a 10 to 40% sucrose gradient (12 ml in Buffer A) with a Pasteur pipette. The gradient was centrifuged for 2 hours and 15 min (36,000 rpm, 4°C) in an SW41 rotor. To fractionate the gradient, the bottom
of the tube (Beckman Ultra-Clear Tube #44059) was punctured by a Tube Piercer (ISCO) and 50% sucrose was pumped into the bottom at 40% speed (1 ml/min) (Tris-Pump, ISCO). Twelve 1 ml fractions were collected from the top of the gradient using the Foxy Jr. Fraction Collector (ISCO) while the OD$_{254}$ was detected and recorded by the UA-6 Detector (ISCO) with chart recorder speed set at 150 and sensitivity at 0.5.

For high salt treatment, the supernatant from the low speed spin (14,000 x g, 4°C) was adjusted to a final concentration of 800 mM KCl and centrifuged through a 10-40% sucrose gradient containing 800 mM KCl. For EDTA treatment, the Mg$^{2+}$ concentration of all solutions was reduced to 0.5 mM, and the supernatant and gradient were adjusted to a final concentration of 50 mM EDTA. The UA-6 Detector sensitivity was set at 1.0.

**RNA extraction from sucrose fractions**

To isolate RNA, 500 μl of each sucrose fraction were extracted with one volume of phenol and subsequently with an equal volume of 1:1 TE-saturated phenol:chloroform (volume:volume) [29]. Nucleic acid was precipitated from the upper aqueous phase by adding 1/10 volume sodium acetate pH 3.5, 2.5 volumes 100% ethanol, and 20 μg glycogen at -20°C [29]. After washing with 1 ml of 70% ethanol, samples were resuspended in RNase-free water and treated with 2 μl RQ1 DNase (2 U) (Promega) in RQ1 DNase buffer for 1 hour at 37°C. The DNase reaction was stopped by extraction with an equal volume of 1:1 TE-saturated phenol:chloroform (volume:volume). The upper aqueous phase was extracted with one volume of chloroform. RNA was precipitated from the
aqueous phase with 1/10 volume sodium acetate, pH 3.5, 2.5 volumes 100% ethanol, and 20 μg glycogen at -20°C. After washing with 1 ml of 70% ethanol, RNA was stored at -80°C as an ethanol precipitate or dissolved in 10 μl RNase-free TE buffer (Ambion). When extracting RNA from high salt or EDTA-containing sucrose fractions, 300 μl of each fraction were first diluted with 300 μl RNase-free water. Also, the final RNA pellet dissolved in 6 μl RNase-free TE buffer. For RNA extractions from EDTA-containing fractions, the sodium acetate pH was increased to 7.0 with 1 M NaOH.

RESULTS

Expression of reporter enzymes in Leishmania tropica

To test the effect of 5’ UTRs on protein expression in Leishmania, eight 12-nucleotide sequences [12] were tested in the context of a 250-nucleotide L. major DHFR/TS intergenic region [13] (Figure 2.1, Table 2.1). These sequences, occurring just upstream of the translation initiation codon, differed only in the -3 to -1 sequence defined as the AUG-proximal region (APR) (Figure 2.1A).

Protein expression was determined by measuring activity of a reporter enzyme glycosylphosphatidylinositol-specific phospholipase C (GPI-PLC), which is not detectable in Leishmania [20]. Each 12-mer was linked to the GPI-PLC coding region and cloned into pX63NEO [13], an expression plasmid that is maintained extrachromosomally in Leishmania at a copy number of approximately 20 per cell (Figure 2.1B) [30]. Stable transfectants were selected, and uncloned cells were used in GPI-PLCp enzyme assays.
Reporter enzyme activity varied significantly when linked to different AUG-proximal regions (Figure 2.2). For example, lacZ-CCC resulted in 10-fold more protein expression than a wild type *T. brucei*-GTA (*Tb*-GTA) 5’ UTR. Strikingly, when lacZ-CCC is compared to lacZ-CTC, the difference in protein expression levels is 700-fold. *Tb*-GTA and lacZ-GTA resulted in similar levels of reporter protein. Six 5’ UTRs (lacZ-CCC, lacZ-GCC, lacZ-CTA, lacZ-ACC, lacZ-ATC, lacZ-CCA) resulted in protein expression at least 4-fold above the baseline observed with *Tb*-GTA and lacZ-GTA. One 5’ UTR, lacZ-CTC, resulted in expression 60-fold below this baseline.

To determine whether the observations made with *GPI-PLC* could be extended to other genes, the effect of the APRs on a second reporter gene, firefly luciferase (*LUC*) was tested [25]. The *LUC* coding region was placed downstream of the AUG-proximal region in pX63NEO (Figure 2.1C), and the activity of LUCp in stable transfectants of *L. tropica* measured [16, 21]. Luciferase activity was affected greatly by the different APRs (Figure 2.3). Two sequences, lacZ-GTA and lacZ-ATC, resulted in relatively low LUCp expression, while lacZ-CTA produced at least 6-fold higher protein expression. Remarkably, four 5’ UTRs (lacZ-CTC, lacZ-GCC, lacZ-CCC, lacZ-ACC) produced at least 2700-fold more protein than lacZ-GTA and lacZ-ATC. These results confirm our earlier data (Figure 2.2), which indicate that AUG-proximal regions of 5’ UTRs can affect protein expression in *Leishmania*.

In most cases, APRs had similar effects on both reporter genes. LacZ-ACC, lacZ-CCC, lacZ-GCC and lacZ-CCA resulted in high expression of GPI-
PLCp and LUCp, and lacz-GTA yielded low amounts of both proteins (Figures 2.2 and 2.3). However, a few APRs gave different results. For example, lacz-ATC gave high GPI-PLCp but low LUCp expression, while the opposite was true of lacz-CTC (low GPI-PLCp and high LUCp expression). We conclude that (i) the coding sequence can sometimes affect the amount of protein expressed with an APR, and (ii) effects of some APRs could be influenced by coding sequence.

We were curious to know if single nucleotide changes in the APR could alter protein expression. The lacz-ATC APR varies from lacz-CTC by only one nucleotide at -3, yet the APRs result in over a 390-fold difference in GPI-PLCp expression (Figure 2.2). Similarly, expression of LUCp is 800-fold higher with lacz-CTC than with lacz-CTA, although these APRs differ only by one nucleotide at -1 (Figure 2.3). Lacz-ACC and lacz-ATC sequences are different only at -2, but have a 3500-fold difference in LUCp expression (Figure 2.3). These data indicate that single nucleotide changes in any position of the AUG-proximal region can affect protein expression.

**Steady-state levels of GPI-PLC and LUC mRNA**

Differences in reporter gene expression might be accounted for by changes in the amounts of mature mRNA. To test this hypothesis, steady state levels of reporter transcripts were measured by Northern blot analysis (Figures 2.4 and 2.5).

Quantities of reporter mRNA did not correlate with protein levels. For example, with *GPI-PLC* as the reporter gene, lacz-GTA and lacz-CCA result in the same level of steady state mRNA, but expression of GPI-PLCp is four times
higher with lacz-CCA as compared to lacz-GTA (Figures 2.2 and 2.4). The discrepancy between mRNA and protein levels was most evident with lacz-CTC produced the highest level of mRNA, but yielded no detectable protein (Figures 2.2 and 2.4). The levels of GPI-PLC mRNA varied 4-fold (Figure 2.4), while the difference in GPI-PLCp expression was 700-fold (Figure 2.3).

Similar results were obtained when LUC was utilized as a reporter gene (Figure 2.5). For example, lacz-ATC and lacz-GCC resulted in similar levels of LUC mRNA, but the LUCp expression with lacz-GCC is three orders of magnitude higher than that observed with lacz-ATC. The amount of LUCp activity produced by the APRs differed 6000-fold, yet the variation in steady state LUC mRNA was only 5-fold.

The data from GPI-PLC and LUC Northern analysis indicate that steady state levels of reporter mRNA do not directly correlate with the amount of reporter protein.

**Trans-splicing and polyadenylation of GPI-PLC pre-mRNA**

Protein-coding genes are transcribed as a part of polycistronic units in trypanosomatids [1, 2]. Messenger RNAs are matured by trans-splicing of the spliced leader (SL) and polyadenylation [3]. Trans-splicing is an intermolecular reaction that adds a conserved 39-41 nucleotide spliced leader (SL) to the 5’ UTR of all nuclear-encoded pre-mRNAs [31, 32]. SL RNA is encoded on a different chromosome than the pre-mRNA.

Although SL is required for the expression of reporter genes in trypanosomatids [16, 21, 33, 34], there is no direct evidence for an effect of the
spliced leader on translation in *Leishmania*. Nevertheless, one possible explanation for our results is that trans-splicing does not occur the pre-mRNAs containing APRs that result in low protein expression (e.g., lacz-CTC and lacz-GTA) (Figure 2.2).

Accordingly, RT-PCR was used to detect the presence of SL on reporter transcripts. A forward primer to the SL sequence from *Leishmania* was used with a reverse primer against the GPI-PLC coding sequence. Only if the SL sequence was contiguous with GPI-PLC mRNA could a product be synthesized in the PCR step. RT-PCR with total RNA from *Leishmania* expressing GPI-PLC resulted in a product of approximately 350 base pairs (Figure 2.6, lanes 1-6). A reaction with RNA from *Leishmania* stably transfected with the plasmid pX63NEO alone did not result in this product (Figure 2.6, lane 7). We conclude that all reporter transcripts were trans-spliced.

In vertebrates, the interaction of the poly(A) tail at the 3' end of an mRNA with the 5' methyl-7-guanosine cap can stimulate translation [35, 36]. It is not known whether a similar mechanism operates in trypanosomatids [37], but a lack of polyadenylation to reporter transcripts might explain the low protein expression observed with some APRs. To test this hypothesis, polyadenylated and non-polyadenylated RNA were fractionated, and Northern dot blot analysis was performed to detect GPI-PLC mRNA (Figure 2.7). A probe to ribosomal protein S8 (rpS8) mRNA, which is found in poly(A)+ RNA, was used as a control [38]. GPI-PLC mRNA was found in the polyadenylated fraction for all transfectants. GPI-PLC mRNA was not found in the poly(A)+ fraction from *Leishmania*.
transfected with pX63NEO plasmid alone. We conclude that reporter transcripts possess poly(A) tails.

Together, RT-PCR and Northern dot blot analysis suggest that defective pre-mRNA processing is not responsible for low levels of protein expression obtained with some APRs.

**Distribution of reporter mRNA between cytosol and ribosomes**

During translation initiation, the 43S ribosomal initiation complex binds to mRNA to form a 48S complex [39]. The 60S large ribosomal subunit joins the 48S complex at the initiator AUG, forming the 80S ribosome or monosome [39]. Polysomes form when multiple ribosomes are recruited to mRNA before translation terminates. Since GPI-PLC mRNA levels did not always correlate with the amount of GPI-PLCp (Figure 2.4), we hypothesized that APRs that lead to high protein expression (e.g., lacz-CCC) recruit ribosomes efficiently (Figure 2.8). To test this idea, the distribution of reporter mRNA between cytosol and ribosomes was investigated *in vivo*.

Four APRs that result in varying levels of GPI-PLCp expression were studied: lacz-CCC (very high), lacz-CTA (high), lacz-GTA (low), and lacz-CTC (very low) (Figure 2.2). Cells stably expressing the GPI-PLC constructs were lysed in detergent, and the supernatant from a low speed spin was separated on a 10-40% sucrose gradient [28]. The gradient was fractionated, and RNA was extracted from each fraction for analysis. The locations of 40S ribosomal subunits, 60S ribosomal subunits, 80S monosomes, and polysomes were
estimated using the OD$_{254}$ absorbance profile and by blotting for ribosomal RNAs (Figure 2.9 A, 2.9 C, and 2.10).

Polysome analysis has not been established for *Leishmania*. Therefore we (i) included tests to determine the distribution of an endogenous mRNA, and (ii) to verify the location of the ribosomes in sucrose gradients. We examined the transcript for a highly abundant *Leishmania* cell surface protein, gp63. Most (69-80%) of the gp63 mRNA was associated with polysome fractions in all transfectants (Figure 2.9 A and 2.9 B). Large subunit ribosomal RNA (rRNA) was detected primarily in the 60S/80S (29-35%) and polysome fractions (48-62%) for all transfectants. Distribution of small subunit rRNA (data not shown) was also in agreement with the OD$_{254}$ absorbance profile.

For APRs that led to high GPI-PLCp expression, lacZ-CCC and lacZ-CTA, (Figure 2.2) most of the GPI-PLC mRNA associated with polysomes (54-61%) (Figure 2.10). Conversely, with lacZ-GTA, which produced low amounts of GPI-PLCp (Figure 2.2), the distribution of GPI-PLC mRNA differed significantly from the others; 42% of the total GPI-PLC mRNA was found in the cytosol, as compared to only 10-16% observed with other 5' UTRs (Figure 2.10). We conclude that the efficiencies of protein synthesis directed by lacZ-CCC, lacZ-CTA, and lacZ-GTA are consistent with their relative ability to recruit ribosomes.

For lacZ-CTC, the distribution of reporter mRNA was unexpected. Sixty percent of GPI-PLC mRNA was bound to polysomes (Figure 2.10), although the APR produced little GPI-PLCp (Figure 2.2). We conclude that a mechanism of control other than ribosome recruitment may operate with lacZ-CTC. In C.
and *Drosophila*, some translationally repressed mRNAs have also been found on polysomes [40, 41].

**Unusual stability of *Leishmania* ribosomes**

We tested the effects of high salt and EDTA on the distribution of the reporter transcript. In most eukaryotes, high salt (800 mM KCl) dissociates non-translating ribosomes [29, 42]. Similarly, 50 mM EDTA disrupts 80S monosomes by chelating Mg\(^{2+}\) required for 40S and 60S ribosomal subunit association [29, 42]. It is expected that successful disruption of ribosomes will result in shifting of the OD\(_{254}\) peaks and rRNA from the heavier fractions of the gradient to the lighter fractions.

When the *Leishmania* lysate was treated with 800 mM KCl before separation on a 10-40% sucrose gradient containing 800 mM KCl, an OD\(_{254}\) peak shifted from fractions 5 and 6 to fractions 3 and 4 (Figure 2.11 A and C). Surprisingly, the movement of the OD\(_{254}\) peak was not reflected in the distribution of either the small subunit rRNA or large subunit rRNA (Figure 2.11 B and D). Similar results were obtained with 50 mM EDTA treatment (data not shown). These data indicate that *Leishmania* ribosomes are resistant to high salt and EDTA. The shift in the OD\(_{254}\) peak could be associated with disruption of macromolecular complexes that are present in larger amounts than ribosomes.

**DISCUSSION**

Trypanosomatids rely mainly on post-transcriptional control mechanisms for gene expression. Untranslated regions of mRNA may influence this type of
regulation, and we are interested in the contribution of 5' UTRs to mRNA translatability in *Leishmania*.

**AUG-proximal regions of 5' UTRs affect protein expression in *Leishmania***

In the context of an endogenous *Leishmania* 5' UTR, we show that AUG-proximal sequences can dramatically affect protein expression in the parasite. APRs tested with a *GPI-PLC* reporter yielded a 700-fold difference in protein expression (Figure 2.2), and with *LUC*, a 6000-fold difference was observed (Figure 2.3). To our knowledge, the magnitude of effects of APRs on protein synthesis is unprecedented in any biological system. Changing nucleotides in the APR yielded 20-fold difference in vertebrate cells and only 2-fold differences in yeast.

We identified three APRs that enhance translation of both reporters: lacz-ACC, lacz-CCC, and lacz-GCC. These sequences may be useful for efficient translation of proteins in *Leishmania*.

Single nucleotide changes in the APR resulted in significant changes in protein expression. For example, lacz-ACC produced 3500-fold more LUCp than lacz-ATC, which differs by only one nucleotide at position -2 (Figure 2.3). Choice of reporter gene sometimes affected the amount of protein obtained. For example, lacz-ATC yielded high expression of GPI-PLCp, but when placed upstream of *LUC*, it gave low protein expression.

**Comparison of effects of AUG-proximal sequences on translation in *Leishmania* and vertebrates**
In vertebrate cells, the Kozak box (GCCA/GCCAUGG) is one “favorable” context for translation initiation [43]. From detailed analysis of this sequence, it is held that A or G should be present at -3 in order for a start codon to be utilized optimally by a ribosome [44, 45]. Similarly, ACC and GCC (with A or G at -3) enhanced protein synthesis in *Leishmania* (Figures 2.2 and 2.3).

In contrast to vertebrate systems, a C at -3 does not result in low protein expression in *Leishmania*. For example, lacz-CCC and lacz-CTA 5’ UTRs yielded amounts of LUCp that were 2900 and 6-fold higher than the baseline, respectively (Figures 2.3). Additionally, a G at -3 does not always result in high reporter protein as illustrated by lacz-GTA and *Tb*-GTA (Figure 2.2). Therefore, the established rules for vertebrate translation may not be directly transferable to trypanosomatid systems.

**APRs control translation post-transcriptionally**

Steady-state levels of reporter mRNA did not specify reporter protein abundance in *Leishmania* (Figures 2.4 and 2.5). The difference in the amount GPI-PLC mRNA from the APRs was 4-fold, but the expression of GPI-PLCp varied 700-fold. APRs gave rise to up to a 6000-fold difference in LUCp expression, yet only a 5-fold difference in steady state LUC mRNA was observed. These data indicate that the effects of the AUG-proximal regions on reporter protein expression are predominantly post-transcriptional.

Although our primary interest was how the 5’ UTRs might influence mRNA translatability, we first examined whether a lack of pre-mRNA processing could explain the results. There is no direct experimental evidence for a role of the
spliced leader or the poly(A) tail [37] in trypanosomatid translation, so we tested for their presence qualitatively. GPI-PLC transcripts from each 5' UTR receive SL and are polyadenylated (Figures 2.6 and 2.7). Therefore, defective pre-mRNA processing was not likely to be the cause of low translation detected with some AUG-proximal regions (e.g., Tb-GTA, lacz-GTA, and lacz-CTC) (Figure 2.2).

**AUG-proximal sequences in 5' UTRs influence ribosome recruitment**

Since the amount of reporter protein did not correlate with abundance of reporter mRNA (Figure 2.4), we checked whether the AUG-proximal regions modulated recruitment of ribosomes (Figure 2.8). Two APRs that resulted in high GPI-PLCp expression, lacz-CCC and lacz-CTA, caused the reporter transcript to associate primarily with polysomes (Figure 2.10). Lacz-GTA, which yielded low GPI-PLCp expression, was ineffective at recruiting ribosomes since 4-fold more reporter mRNA remained in the cytosol as compared to the other APRs tested (Figure 2.10). From these data we conclude that efficiency of ribosome recruitment correlates, in general, with translation competency of the mRNA.

While the effects of lacz-CCC, lacz-CTA, and lacz-GTA on reporter mRNA distribution between cytosol and ribosomes could be explained by a ribosome recruitment model (Figures 2.8 and 2.10), the effect of lacz-CTC could not be accounted for by that mechanism. Lacz-CTC produced low amounts of GPI-PLCp (Figure 2.2), but the GPI-PLC mRNA was associated with polysomes (Figure 2.10). We conclude that a post-recruitment mechanism may block translation of lacz-CTC GPI-PLC mRNA on ribosomes associated with lacz-CTC
GPI-PLC mRNA. Although unusual, this phenomenon is not without precedent. In *C. elegans*, *lin-14* mRNA is translationally repressed, yet the transcript is associated with polysomes [40]. Similarly, protein synthesis from nanos (*nos*) mRNA is blocked by the binding of Smaug protein to its 3’ UTR, but the transcript remains bound to polysomes [41]. In both studies it was suggested that inhibition of protein synthesis occurs after translation initiation [40, 41].

**Unusual stability of *Leishmania* ribosomes in high salt and EDTA**

We attempted to probe the nature of the interactions between lacZ-CTC GPI-PLC mRNA and ribosomes. Both high salt and EDTA disrupt polysomes in vertebrates and yeast [29, 42]. Surprisingly, *Leishmania* ribosomes were not dissociated by high salt (800 mM KCl) (Figure 2.11) or 50 mM EDTA (data not shown). Interestingly, although OD$_{254}$ absorbance peaks from the sucrose gradient shifted upon exposure to either high salt or EDTA (Figure 2.11 A and C), the distribution of small and large subunit rRNA across the gradient did not change (Figure 2.11 B and D). This observation indicates that other macromolecules (with absorbance at OD$_{254}$) were affected by these treatments. It is likely that other ribonucleoprotein complexes (such as spliceosomes), contribute disproportionately to the OD$_{254}$ absorbance. These complexes may be shifting to the lighter portion of the sucrose gradient following high salt or EDTA treatment.

In summary, we report the large effects of AUG-proximal regions on protein expression in *Leishmania*. APRs can modulate translation by ribosome recruitment, and by other unidentified mechanisms. These and other
investigations of the role of 5' UTRs in protein synthesis will increase our understanding of post-transcriptional control of gene expression in *Leishmania*.
REFERENCES


Table 2.1: Sequences near the translation initiation codon in the 5’ untranslated region (5’ UTR). The -12 to -1 regions of the 5’ UTR contain three sequences: (i) a polypurine hexamer (bold italics), (ii) an *E. coli lacZ* gene spacer (underlined), and (iii) nucleotides of the AUG-proximal region (APR) at -3, -2 and -1 (bold). 5’ UTRs are named according to the spacer and APR. For example, *Tb*-GTA refers to a *T. brucei GPI-PLC* 5’ UTR with GTA at -3 to -1.
Table 2.1

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>lacz-ACC</td>
<td>AGGAGGCAGACCatg</td>
</tr>
<tr>
<td>lacz-ATC</td>
<td>AGGAGGCAGATCatg</td>
</tr>
<tr>
<td>lacz-CCA</td>
<td>AGGAGGCAGCCAtg</td>
</tr>
<tr>
<td>lacz-CCC</td>
<td>AGGAGGCAGCCCatg</td>
</tr>
<tr>
<td>lacz-CTA</td>
<td>AGGAGGCAGCTAatg</td>
</tr>
<tr>
<td>lacz-CTC</td>
<td>AGGAGGCAGCTCatg</td>
</tr>
<tr>
<td>lacz-GCC</td>
<td>AGGAGGCAGGCCatg</td>
</tr>
<tr>
<td>lacz-GTA</td>
<td>AGGAGGCAGGTAatg</td>
</tr>
<tr>
<td>Tb-GTA</td>
<td>GTTAAGAATCATTTGTAatg</td>
</tr>
</tbody>
</table>
Figure 2.1: AUG-proximal region (APR) and expression plasmids for *Leishmania*.  

**A)** The 5’ UTR is defined as the region between the splice leader acceptor site and the translation initiation codon (AUG). The AUG-proximal region (APR) is the sequence at -3, -2, and -1. SL = spliced leader, DHFR/TS = dihydrofolate reductase/thymidylate synthetase, GPI-PLC = glycosyl-phosphatidylinositol-specific phospholipase C, LUC = luciferase.  

**B)** The -12 to -1 sequences (Table 2.1) were placed upstream of a GPI-PLC coding region, and cloned into a Bam H I site of pX63NEO.  

**C)** Plasmid pX63NEO-LUC is identical to pX63NEO-GPI-PLC except that the GPI-PLC coding region has been replaced with LUC. DHFR/TS intergenic regions (upstream and downstream sequences (US and DS)) from *Leishmania* are required for pre-mRNA processing of NEO and GPI-PLC.
Figure 2.1

A

B

C
Table 2.2: Luciferase pX63NEO Primers. The primers were used to amplify luciferase fragments from p220S. Amplified DNAs were cloned into pX63NEO (Figure 1C). aa = amino acids.
<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>3'-LUC-BamHI</td>
<td>CGCGGATCCttacaatggactttcgcc</td>
<td><em>Bam</em> HI site (bold), nucleotides complementary to luciferase sequence encoding aa 545-551 (lower case)</td>
</tr>
<tr>
<td>5' <em>Bam</em>HI-lacz-ACC-LUC</td>
<td>CGCGGATCCTTAACACAGGAGGCAGACCatggaagcgccccaaacata</td>
<td><em>Bam</em> HI site (bold), purine hexamer (underline), lacz spacer (outline), APR (italics, bold), luciferase sequence encoding aa 1-6 (lower case)</td>
</tr>
<tr>
<td>5' <em>Bam</em>HI-lacz-ATC-LUC</td>
<td>CGCGGATCCTTAACACAGGAGGCAGATCatggaagcgccccaaacata</td>
<td>See 5' <em>Bam</em>HI-lacz-ACC-LUC features</td>
</tr>
<tr>
<td>5' <em>Bam</em>HI-lacz-CCC-LUC</td>
<td>CGCGGATCCTTAACACAGGAGGCAGCCCatggaagcgccccaaacata</td>
<td>See 5' <em>Bam</em>HI-lacz-ACC-LUC features</td>
</tr>
<tr>
<td>5' <em>Bam</em>HI-lacz-CTA-LUC</td>
<td>CGCGGATCCTTAACACAGGAGGCAGCTAatggaagcgccccaaacata</td>
<td>See 5' <em>Bam</em>HI-lacz-ACC-LUC features</td>
</tr>
<tr>
<td>5’ <em>Bam</em>HI-lacz-CTC-LUC</td>
<td>CGC&lt;sub&gt;GGATCC&lt;/sub&gt;TAAACACAGGAGG&lt;sub&gt;CAG&lt;/sub&gt;CT&lt;sub&gt;C&lt;/sub&gt;atgaagacgc&lt;sub&gt;c&lt;/sub&gt;aaaaaaaaacata</td>
<td>See 5’ <em>Bam</em>HI-lacz-ACC-LUC features</td>
</tr>
<tr>
<td>-------------------------</td>
<td>------------------------------------------------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>5’ <em>Bam</em>HI-lacz-GCC-LUC</td>
<td>CGC&lt;sub&gt;GGATCC&lt;/sub&gt;TAAACACAGGAGG&lt;sub&gt;CAG&lt;/sub&gt;GCCatgaagacgc&lt;sub&gt;c&lt;/sub&gt;aaaaaaaaacata</td>
<td>See 5’ <em>Bam</em>HI-lacz-ACC-LUC features</td>
</tr>
<tr>
<td>5’ <em>Bam</em>HI-lacz-GTA-LUC</td>
<td>CGC&lt;sub&gt;GGATCC&lt;/sub&gt;TAAACACAGGAGG&lt;sub&gt;CAG&lt;/sub&gt;GTAatgaagacgc&lt;sub&gt;c&lt;/sub&gt;aaaaaaaaacata</td>
<td>See 5’ <em>Bam</em>HI-lacz-ACC-LUC features</td>
</tr>
</tbody>
</table>
Table 2.3: Primers for probe synthesis and RT-PCR.  
A) The primers were used for PCR. The products were labeled with psoralen-biotin and used as probes in Northern analysis.  
B) RT-PCR analysis was performed using the primers. All primer sequences were selected using Laser Gene software (DNASTAR Inc.). aa = amino acids.
### Table 2.3

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Target for DNA Amplification</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>KCR21</td>
<td>A forward primer to <em>GPI-PLC</em> from <em>T. brucei</em>, <em>GPI-PLC</em> sequence encoding aa 1-15</td>
<td>pX63NEO-GPIPLC</td>
<td>TTTGGTGGTGTAAAGTGGTCACCGCAGTCATGGATGAGTGACACG</td>
</tr>
<tr>
<td>KCR8</td>
<td>A reverse primer to <em>GPI-PLC</em> from <em>T. brucei</em>, nucleotides complementary to <em>GPI-PLC</em> sequence encoding aa 132-146</td>
<td>pX63NEO-GPIPLC</td>
<td>TATGTGGATCCTTATTAGAAATCGAATGACAAATTCGTTG GC</td>
</tr>
<tr>
<td>FLtrpS8</td>
<td>A forward primer to ribosomal protein S8 (<em>rpS8</em>) from <em>Leishmania</em>, <em>rpS8</em> sequence encoding aa 1-7</td>
<td><em>Leishmania</em> genomic DNA</td>
<td>CTGCATAAGCGCAAGATCACC</td>
</tr>
<tr>
<td>RLtrpS8</td>
<td>A reverse primer to <em>rpS8</em> from <em>Leishmania</em>, nucleotides complementary to <em>rpS8</em> sequence encoding aa 147-153</td>
<td><em>Leishmania</em> genomic DNA</td>
<td>CTTGTTACGTCGTACTTGTC</td>
</tr>
<tr>
<td>Primer</td>
<td>Description</td>
<td>Sequence</td>
<td>Gene</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>----------</td>
<td>------</td>
</tr>
<tr>
<td><strong>FLUC</strong></td>
<td>A forward primer to LUC from firefly, LUC sequence encoding aa 313-320</td>
<td>p220S</td>
<td>GCTTCTGGGGGCGCACCTCTTTTC</td>
</tr>
<tr>
<td><strong>RLUC</strong></td>
<td>A reverse primer to LUC from firefly, nucleotides complementary to LUC sequence encoding aa 444-452</td>
<td>p220S</td>
<td>AGCGGGGGCCACCTGTATCTTTGTA</td>
</tr>
<tr>
<td><strong>Fgp63</strong></td>
<td>A forward primer to gp63 from Leishmania, gp63 sequence encoding aa 715-724</td>
<td>Leishmania genomic DNA</td>
<td>GACCGATGCAGGCACGCGTCG</td>
</tr>
<tr>
<td><strong>Rgp63</strong></td>
<td>A reverse primer to gp63 from Leishmania, nucleotides complementary to gp63 sequence encoding aa 1043-1052</td>
<td>Leishmania genomic DNA</td>
<td>GAAGTCGGTGTTGCTGAAGCCCTCG</td>
</tr>
<tr>
<td><strong>FgamrLtRNA</strong></td>
<td>A forward primer to the large subunit rRNA (g) from Leishmania¹; nucleotides 715-743</td>
<td>Leishmania genomic DNA</td>
<td>CGCAAGCGCAAATGAAATACCACCACCTCG</td>
</tr>
</tbody>
</table>

¹ In *Leishmania*, the 28S rRNA is fragmented into two large (a and b) and four small molecules (c, d, e and x).
<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Description</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>RgamrLtRNA</td>
<td>A reverse primer to the large subunit rRNA (16S) from <em>Leishmania</em>; complementary to nucleotides 1027-1052</td>
<td>Leishmania genomic DNA; TGCCGCCCCAGCCAAACTCCCCATCT</td>
</tr>
<tr>
<td>F18SrLtRNA</td>
<td>A forward primer to the small subunit rRNA (18S) from <em>Leishmania</em>; nucleotides 1430-1456</td>
<td>Leishmania genomic DNA; CACCGGAAAGCTTTGAGGTTACAGTCT</td>
</tr>
<tr>
<td>R18SrLtRNA</td>
<td>A reverse primer to the small subunit rRNA (18S) from <em>Leishmania</em>; complementary to nucleotides 1714-1741</td>
<td>Leishmania genomic DNA; ACCCGCGATGAGTTGCTATTCTATGG</td>
</tr>
<tr>
<td>KCR2</td>
<td>A reverse primer to <em>GPI-PLC</em> from <em>T. brucei</em>, nucleotides complementary to <em>GPI-PLC</em> sequence encoding aa 52-58</td>
<td>Leishmania total RNA; CTCCCTCCATATGTGACCTTT</td>
</tr>
</tbody>
</table>

**B**
<table>
<thead>
<tr>
<th>LmSL2</th>
<th>A forward primer to nucleotides 1-39 of <em>L. tropica</em> spliced leader sequence (underlined), <em>Bam</em> HI site (bold), <em>Cla</em> I site (outline)</th>
<th><em>Leishmania</em> total RNA</th>
<th>GGATCCATCGATAACTAACGCTATATAAGTA TCAGTTTCGTACTTTATTTG</th>
</tr>
</thead>
</table>


Figure 2.2: Effect of AUG-proximal sequences on GPI-PLC enzyme activity.

Transfected *Leishmania* (10^8 cells) were lysed hypotonically, centrifuged, and the pellet extracted with detergent. The detergent extract was centrifuged, and the supernatant was used in a GPI-PLC assay with [3H]myristate-labeled membrane form variant surface glycoprotein ([3H]mfVSG) as substrate. Total protein in the lysate was determined with a bicinchoninic acid assay (Pierce). Enzyme assays were performed in triplicate. Data from a representative experiment is shown.
Figure 2.2

![Bar chart showing specific activity of GPI-PLC (units/mg) for different 5’ UTR sequences.](image)
**Figure 2.3: Effect of AUG-proximal sequences on luciferase enzyme activity.** Stably transfected *Leishmania* (10^7 cells) were lysed with 0.2% Triton X-100, and the lysate was used in a luciferase assay with luciferin as substrate. Relative light units were measured and corrected for total protein with a bicinchoninic acid assay (Pierce). Enzyme assays were performed in triplicate. Data from a representative experiment is shown.
Figure 2.3

![Bar chart showing luciferase activity in 5' UTR]
**Figure 2.4: Northern analysis of GPI-PLC transcripts.**  

**A)** Total RNA was isolated from $10^8$ *Leishmania* with TriReagent (Sigma). Twenty micrograms of RNA was separated on a 1% formaldehyde agarose gel, and electrobotted to a BrightStar-Plus membrane (Ambion). After UV crosslinking, the membrane was probed with a psoralen-labeled probe for GPI-PLC or a loading control, ribosomal protein S8 (*rpS8*). Probes were detected using BioDetect (Ambion), and then the membrane was exposed to BioMax MR film (Kodak). The graph of the specific enzyme activity of GPI-PLC came from data presented in Figure 2.2.  

**B)** Comparison of normalized GPI-PLC mRNA signal to GPI-PLC activity. GPI-PLC and *rpS8* mRNA signals were quantitated directly from the membrane with a Genegnome Chemiluminescence Imager and Genetools software (Syngene). The signal from GPI-PLC mRNA was divided by that for *rpS8* mRNA to correct for differences in amount of total mRNA loaded per lane.
Figure 2.4

A

![Graph showing specific activity of GPI-PLC](image)

B

![Graph showing specific activity of GPI-PLC](image)
Figure 2.5: Northern analysis of luciferase transcripts. A) Northern analysis was performed as detailed in the legend to Figure 2.4 using probes for luciferase (\textit{LUC}) or ribosomal protein S8 (\textit{rpS8}) (a loading control). Data for the graph of luciferase activity was taken from Figure 2.3. B) Comparison of normalized LUC mRNA signal to luciferase activity. Quantitation and correction of the signal for LUC mRNA was done as described in the legend to Figure 2.4.
Figure 2.5

A

![Graph showing Luciferase Activity (Relative Light Units per µg protein) for various sites: lacZ ACC, ATC, CCC, CTA, CTC, GCC, GTA, pX63 NEO. The graph also includes gel images for LUC and rpS8 with lanes 1 to 8.]

B

![Graph showing a similar comparison of Luciferase Activity (Relative Light Units per µg protein) with additional mRNA signal comparing LUC mRNA signal / rpS8 signal for lacZ-ACC, ATC, CCC, CTA, CTC, GCC, GTA, pX63 NEO.]
Figure 2.6: RT-PCR Analysis of GPI-PLC transcripts. One µg of total RNA from *Leishmania* was used for RT-PCR. cDNA synthesis was directed by KCR2, a reverse primer against the GPI-PLC coding sequence. In the PCR step, KCR2 was used with LmSL2, a forward primer to the spliced leader sequence. The resulting product (approximately 350 bp) was resolved on a 1.2% agarose gel and detected by ethidium bromide staining. Lane “M” contains the molecular weight DNA marker, φX174/Hae III (Promega).
**Figure 2.7: Polyadenylation of GPI-PLC transcripts.** Total RNA was isolated from *Leishmania* expressing GPI-PLC as described in the legend to Figure 2.4. Polyadenylated and non-polyadenylated RNA were separated by adsorbing twice to a Poly(A) Plus matrix. For dot blot analysis, 1 μg of poly(A+) RNA and 1 μg of poly(A-) RNA were spotted onto two Bright Star Plus™ membranes. Crosslinking, hybridization with *GPI-PLC* or *rpS8* psoralen-labeled probes, and detection were performed as described in the legend to Figure 2.4.
Figure 2.8: Hypothetical effects of 5' UTRs on ribosome binding to mRNAs.

A) A 5' UTR is deficient in ribosome recruitment. Reporter mRNA is found in the cytosol and possibly bound to the 40S ribosomal subunit. Little or no association with monosomes or polysomes is observed. B) Inefficient recruitment of ribosomes by a 5' UTR. Reporter transcript is associated with 80S ribosomes, 60S and 40S ribosomal subunits, and cytosol. C) Efficient ribosome recruitment to a 5' UTR. Reporter mRNA is predominantly associated with polysomes.
Figure 2.8

Possible effects of synthetic 5' UTRs on ribosome recruitment

A) Deficient

B) Inefficient

C) Efficient
**Figure 2.9: Distribution of gp63 mRNA and large subunit rRNA on ribosomes (controls).** *Leishmania* expressing GPI-PLC were lysed, and a supernatant from a 14,000 x g centrifugation was separated on a 10-40% sucrose gradient. Twelve 1 ml fractions were collected. RNA was extracted from the fractions, and 1/5 of the RNA was spotted onto a Bright-Star Plus membrane. After UV crosslinking, the membrane was probed with a psoralen-labeled probe as described in the legend to Figure 2.4. **A)** DNA probes were as follows: gp63 or large subunit rRNA ([subunit]). Representative blots are presented. **B)** Quantitation of gp63 mRNA and large subunit rRNA. Signals were quantitated as described in the legend to Figure 2.4. The proportion of the RNA associated with cytosol, 40S ribosomes, 60S/80S ribosomes, and polysomes is presented. Representative graphs are shown.
Figure 2.9

A

Fraction:

1 2 3 4 5 6 7 8 9 10 11 12

gp63 mRNA

large subunit rRNA

cytosol 40S 60S 80S monosomes polysomes

cytosol 60S 80S monosomes polysomes

B

Percent of specified RNA signal

gp63 mRNA
large subunit rRNA

cytosol 40S 60S/80S polysomes
Figure 2.10: Distribution of GPI-PLC mRNA on ribosomes.  A) Separation of ribosomes and detection of GPI-PLC mRNA were done as described in the legends to Figures 2.9 and 2.4.  B) Quantitation of GPI-PLC mRNA signal was performed as described in the legend to Figure 2.4. The proportion of the GPI-PLC mRNA signal associated with cytosol, 40S ribosomes, 60S/80S ribosomes, and polysomes is shown.
Figure 2.10

A

Leishmania cells:
lacz-CCC
lacz-CTC
lacz-CTA
lacz-GTA

GPI-PLC mRNA

B

GPI-PLC mRNA

Percent of total GPI-PLC mRNA signal

Leishmania cells:
lacz-CCC
lacz-CTC
lacz-CTA
lacz-GTA

cytosol 40S 60S/80S polysomes
Figure 2.11: Distribution of rRNA following high salt treatment. Separation of ribosomes was done as described in the legend to Figure 2.9. The lysate was adjusted to a final concentration of 800 mM KCl, and the sucrose gradient was supplemented with 800 mM KCl. Detection of A) large subunit rRNA and B) small subunit rRNA was done as described in the legend to Figure 2.4. Representative blots are shown. Quantitation of C) large subunit rRNA and D) small subunit rRNA signals was performed as detailed in the legend to Figure 2.3. The proportion of signal for rRNA found in association with cytosol, 40S ribosomes, 60S/80S ribosomes, and polysomes is shown. Representative graphs are presented.
Figure 2.11

A. Cytosol

B. Cytosol

C. Large subunit rRNA

D. Small subunit rRNA
CHAPTER 3

NON-RANDOM DISTRIBUTION OF TRI-NUCLEOTIDES IN THE AUG-PROXIMAL REGION OF *LEISHMANIA* GENES

---

1 Stanton, Julie D., Judy H. Willis, and Kojo Mensa-Wilmot. To be submitted to *Journal of Biological Chemistry*
ABSTRACT

Trypanosomatids rely predominantly on post-transcriptional processes for controlling gene expression. In other biological systems, 5’ untranslated regions (5’ UTRs) of mature mRNA contribute immensely to post-transcriptional regulation of gene expression. The AUG-proximal region (APR) (defined as positions -3, -2, and -1 upstream of the translation initiation codon) can dramatically affect protein synthesis in Leishmania. To investigate the potential of APRs to influence gene expression, we have analyzed APRs from 200 Leishmania major genes. Sequence alignment did not produce a convincing APR consensus motif. The difference between the observed distribution of tri-nucleotides in the APR and the distribution expected for random occupancy was statistically highly significant. Of the 64 possible tri-nucleotides, 25% were absent from the data set, at a frequency 6-fold higher than expected. Rare tri-nucleotides included: ATA, ATG, ATT, CAT, CCA, CCC, CCT, CTA, CTT, GGG, GGT, TAA, TAC, TAT, TGT, TTA, TTC, and TTT. The frequencies of some popular tri-nucleotides were up 7.724 x 10^5-fold higher than expected. Frequently used tri-nucleotides were CAC, ACA, ACC, ATC, GCC, GCG, and GTC.

The contribution of di-nucleotide patterns to tri-nucleotide usage was revealed using nearest neighbor analysis. Rare tri-nucleotides belonged to di-nucleotide patterns TXT, TXA, CXT, AXT, XTT, XTG, XTA, XGT, XAT, TTX, TAX, CCX, and GGX (where X = any nucleotide). Frequently used tri-nucleotides were part of the di-nucleotide patterns CXC, GXC, AXC, AXA, XTC,
XCG, XCC, XAC, GCX, and ACX. Some popular motifs were observed at frequencies up to $1.359 \times 10^9$ higher than expected for random occurrence. The difference between the observed and expected distributions di-nucleotide patterns was statistically highly significant. The possible physiological relevance of the unusual distribution of tri-nucleotides in *Leishmania* 5' UTRs is discussed.

\* The abbreviations used are: A, adenine; APR, AUG-proximal region (nucleotides at -3, -2, and -1 upstream of AUG); AUG, translation initiation codon; C, cytosine; G, guanine; T, thymine; UTR, untranslated region
INTRODUCTION

*Leishmania* species are the causative agents of the disease leishmaniasis, which infects over 12 million people worldwide. According to the World Health Organization, ten percent of the world’s population is at risk of contracting the disease, and one to two million new cases are reported each year. Manifestations of leishmaniasis range from self-healing lesions of the skin to fatal hepatosplenomegaly [1].

Control of gene expression in *Leishmania* is primarily post-transcriptional because most promoters are constitutively active [2, 3]. Due to the paucity of transcriptional control, untranslated regions (UTRs) are likely to be important for regulation at the levels of pre-mRNA processing, message stability, transcript localization, and mRNA translatability, as documented in other biological systems [4]. The potential role of *Leishmania* 5' UTRs in controlling gene expression has received little attention. We are interested in the contribution of AUG-proximal regions, which we define as the nucleotides at positions -3, -2, and -1 upstream of the translation initiation codon, to gene expression in *Leishmania*.

APRs have been studied in other organisms. For vertebrates, a consensus sequence of G\(^{-6}\)C\(^{-5}\)C\(^{-4}\)R\(^{-3}\)C\(^{-2}\)C\(^{-1}\)augG\(^{+4}\) (where R is a purine) was determined by an alignment of 699 genes [5]. In experimental studies, changing the nucleotides in the APR resulted in up to a 20-fold difference in protein synthesis *in vivo* [6]. Alignment of 96 *Saccharomyces cerevisiae* genes produced a consensus sequence of A\(^{-3}\)A\(^{-2}\)A\(^{-1}\) [7]. Modification of this APR yielded less than a two-fold difference in reporter protein synthesis *in vivo* [8].
In *Leishmania*, it is not known whether there is an APR consensus sequence. However, APRs can have a tremendous impact on protein expression (Chapter Two, Figures 2.2 and 2.3). We tested eight synthetic 12-nucleotide sequences [9] in the context of an 180-nucleotide natural *Leishmania* 5’ UTR [10]. Changing the nucleotides in the APR yielded up to a 6000-fold difference in reporter protein expression (Chapter Two, Figures 2.2 and 2.3). The quantity of reporter protein did not correlate with the amount reporter mRNA (Chapter Two, Figures 2.4 and 2.5). In most cases, the effect of the APRs on translation reflected the efficiency of ribosome recruitment to the transcripts (Chapter Two, Figure 2.10).

Since APRs can dramatically affect protein synthesis in *Leishmania*, we hypothesized that these sequences might not be represented randomly in the genome. To test this idea, we studied tri-nucleotides in the APR of 200 5’ UTRs from *Leishmania major*. However, an informative APR consensus sequence was not found from alignment of the 5’ UTRs. The distribution of tri-nucleotides in this region is not random; some tri-nucleotides are frequently used and others are rare. Nearest neighbor analysis showed that frequently used and rare tri-nucleotides belong to specific patterns of di-nucleotides. We discuss a possibility that di-nucleotide motifs in the APR of *Leishmania* have biological consequences for the parasite.
METHODS

Genomic sequence analysis

DNA sequences were obtained from the Sanger Centre *Leishmania major* genome project (http://www.sanger.ac.uk/Projects/L_major/). The 200 genes used in this study are found on chromosomes 1, 3, 4, 5, 19 and 23 (Appendix). To avoid uncertainty about assignment of translation initiation codons, only 5’ UTRs from coding sequences for proteins with homologs in other organisms were included in the data set.

To obtain an APR consensus sequence from the 200 5’ UTRs, the frequency that each nucleotide (A, C, G, or T) occupied positions -3, -2 and -1 was calculated. Each nucleotide has a 25% chance of occupying a position if selection is random. A consensus nucleotide was assigned if the frequency of that nucleotide was at least twice that amount (50% or more) [11].

**Tri-nucleotide occupancy of the AUG-proximal region: theoretical considerations**

As defined for this study, the AUG-proximal region (APR) contains tri-nucleotides. Any of four nucleotides (A, C, G, or T) can independently occupy the three positions of the APR. Thus the total number of possible tri-nucleotide sequences is $4^3$ or 64. Therefore, the probability ($p$) of one of the 64 possible tri-nucleotides occupying the APR is 1/64.

The binomial equation ($P(x) = n!/(x!(n-x)!) \ p^x(1-p)^{n-x}$) (where $p =$ probability, $n =$ number of independent trials, $x =$ success) was used to determine if the distribution of tri-nucleotide sequences in the 200 5’ UTRs was random. The expected frequency $P(x)$ for observing a tri-nucleotide a specific number of
times \((i.e.,\) success) \((x)\) in 200 independent trials \((n)\) with a probability \((p)\) of \(1/64\) was calculated. The expected and observed frequencies for success between 0 and 17 were calculated; the maximum number of times a tri-nucleotide was observed was 17. The chi-square equation was used to determine if the observed and expected distributions of tri-nucleotides were different. The difference between distributions was statistically significant if \(P < 0.01\).

“Nearest neighbor” analysis

The 64 possible tri-nucleotides were grouped into three sets of 16 di-nucleotide patterns: (i) di-nucleotides at -3 and -1 \((e.g.\ A^{-3}X^{-2}A^{-1})\) \((where\ X = any\ nucleotide)\), (ii) di-nucleotides at -2 and -1 \((e.g.\ X^{-3}A^{-2}A^{-1})\), and (iii) di-nucleotides at -3 and -2 \((e.g.\ A^{-3}A^{-2}X^{-1})\). Each pattern has four members. For example, the di-nucleotide pattern \(A^{-3}A^{-2}X^{-1}\) consists of AAA, AAC, AAG and AAT.

The binomial equation \((p = 1/16, n = 200)\) was used to compare the distribution of expected and observed di-nucleotide frequencies, as described for tri-nucleotides above.

RESULTS

APR consensus sequence for Leishmania 5’ UTRs

We attempted to derive a consensus sequence for the APR by aligning 200 Leishmania 5’ UTRs. If selection of each nucleotide is random, then on average a nucleotide has a 25% chance of being present at a given position. A consensus nucleotide must dominate a position. Therefore, the nucleotide must constitute at least 50% of the population [11]. Using this criterion, a degenerate
consensus sequence of $X^3X^2C^{-1}$ (where $X$ = any nucleotide) was obtained (Figure 3.1).

**Sequences in the APR for 200 *Leishmania* genes**

The APR consensus sequence provided little predictive information, but experimentally, APRs can greatly influence protein expression in *Leishmania* (Chapter Two, Figures 2.2 and 2.3). Therefore, we abandoned the consensus approach to analyzing the data in favor of studying the actual sequences in the APRs.

In the APR, the total number of possible tri-nucleotides is $4^3$ or 64. Only 48 of the possible tri-nucleotides were observed in the data set (Figure 3.2). The maximum number of times a tri-nucleotide occurred was 17 (GCC). Frequently used tri-nucleotides were defined as those present at least half of the maximum observed number (eight times or more). In the data set, frequently used tri-nucleotides were: ACA, ACC, ATC, CAC, GCC, GCG, and GTC. Absent from the data set were eighteen tri-nucleotides: ATA, ATG, ATT, CAT, CCA, CCC, CCT, CTA, CTT, GGG, GGT, TAA, TAC, TAT, TGT, TTA, TTC, and TTT.

To test whether the observed frequencies of tri-nucleotides were consistent with that expected from random distribution, we applied the binomial equation\(^1\) to the data set. The probability ($p$) of a tri-nucleotide occupying an APR is $1/64$. The expected frequency ($P(x)$) for observing a tri-nucleotide a specified number of times (i.e., success) in 200 independent trials ($n$) was calculated. The expected and observed frequencies for success ($x$) between 0

\[ P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \]

\(^1\)
and 17 were compared (Figure 3.3A). (The maximum number of times a tri-nucleotide was observed was 17, although up to 200 is possible.)

The proportion of non-represented tri-nucleotides (x = 0) was over six times higher than the expected frequency if selection was random (Figure 3.3A). It was expected that most tri-nucleotides would be present between one and five times (Figure 3.3). However, some tri-nucleotides were observed 11 times (ACA), 12 times (ACC), 15 times (ATC), and 17 times (GCC), at frequencies up to $7.724 \times 10^5$ times higher than the expected (Figure 3.3B). Using chi-square analysis, the difference between the distributions of expected and observed frequencies was found to be statistically highly significant (P < 0.01).

**Nearest neighbor analysis of the APR**

To learn the “rules” that might govern tri-nucleotide occupancy of *Leishmania* APRs, nearest neighbor analysis was applied to the data set. The rationale for this approach is based on our experimental data, which indicate that the effects of APRs on translation cannot be explained by single nucleotide changes alone. For example, reporter protein expression was seven times higher with CTA than with GTA (Chapter Two, Figure 2.2). One possible interpretation of the data is that a G at -3 is inhibitory. However a G at -3 did not generally result in low protein expression; GCC enhanced expression as compared to GTA (Chapter Two, Figure 2.2).

A di-nucleotide pattern is the next simplest motif that could explain how APRs might influence translation in *Leishmania*. The 64 possible tri-nucleotides were grouped into three sets consisting of (i) di-nucleotides at -3 and -1, (e.g. A-
and (iii) di-nucleotides at -3 and -2 (e.g. A\(^{-3}\)X\(^{-2}\)A) (where X is any nucleotide) (Figure 3.4). Each set is comprised of 16 possible di-nucleotide patterns. For example, pattern A\(^{-3}\)X\(^{-2}\)A includes the tri-nucleotides AAA, ACA, AGC, and ATC.

The AXC di-nucleotide pattern was used 40 times, the maximum number observed. Frequently used di-nucleotide patterns present at least half the maximum number (20 times or more) were CXC, GXC, AXC, AXA, XTC, XCG, XCC, XAC, GCX, and ACX (Table 3.1). Di-nucleotide patterns observed 4 times (10% of the maximum number) or less were classified as “rare”. Rare di-nucleotide patterns were: TXT, TXA, CXT, AXT, XTT, XTG, XTA, XGT, XAT, TTX, TAX, CCX, and GGX (Table 3.1).

The effect of base content on the use of di-nucleotide patterns was examined. Eleven of the thirteen rare di-nucleotide motifs have T in at least one position, and five contain an A (Table 3.1). However, five least-used motifs contain G or C, and two of these contain G or C exclusively (e.g., GGX and CCX). Therefore, rare patterns are not strictly AT-rich. Nevertheless, it is unusual to find two Ts in an APR (Table 3.1) Nine of the ten frequently used di-nucleotide motifs contain a C (Table 3.1). Both G and A are found in three popular patterns each, but T is only found in one (Table 3.1). We conclude that frequently used patterns are generally C-rich and T-deficient.

We also considered the contribution of nucleotide positioning within the APR to di-nucleotide pattern usage. The rare motif CXT differs from the
frequently used motif XTC by the positioning of -3/-1 nucleotides and -1/-2 nucleotides, respectively (Table 3.1). Similarly, XCC was frequently used, yet CCX was rare. We conclude that both the composition and sequence of nucleotides within the APR are important for establishing the frequency with which a di-nucleotide motif was observed.

To determine whether di-nucleotide pattern usage in the APR is random, we applied the binomial equation \( p = 1/16, n = 200 \) to the data set. The expected and observed di-nucleotide frequencies were compared for each of the three sets of patterns (-3/-1, -2/-1, -3/-2) (Figure 3.5). Most of the patterns were expected to be present between nine and 16 times. Strikingly, popular motifs were used up to 40 times, at a frequency \( 1.359 \times 10^9 \)-fold more than expected. For each of the three sets, the difference between the distributions of the expected and observed di-nucleotide pattern frequencies was statistically highly significant by chi-square analysis (\( P < 0.01 \)).

**DISCUSSION**

In *Leishmania*, changing the nucleotides in the AUG-proximal region of a 5’ UTR can greatly affect translation (Chapter Two, Figures 2.2 and 2.3). As a result, we hypothesized that APRs are functionally significant, and therefore, presence of tri-nucleotides in this region might not be random. To assess these theories, we analyzed tri-nucleotides in the APRs of 200 *Leishmania* 5’ UTRs. **Frequency of tri-nucleotides in the AUG-proximal region of Leishmania**
From examination of the sequences, several conclusions can be made. First, some tri-nucleotides are rarely found in the APR. Twenty-five percent of the 64 possible tri-nucleotides were absent from the data set (Figure 3.2). The frequency of non-used tri-nucleotides was six times higher than expected for random occurrence (Figure 3.3B). Second, some tri-nucleotides were over-represented in the APR. Sequences such as ACA, ACC, ATC, CAC, GCC, GCG, and GTC were observed at frequencies between ten and $7.724 \times 10^5$ times higher than expected (Figures 3.2 and 3.3A). The difference between the distributions of observed and expected tri-nucleotide frequencies was statistically highly significant ($P < 0.01$) (Figure 3.3). From these data we conclude that tri-nucleotides in the APR of *Leishmania* 5’ UTRs are not random.

**Di-nucleotide patterns in the APR**

To determine the possible guidelines for occupancy of the APR, we first attempted to assign a consensus sequence from an alignment of the 200 5’ UTRs. The resulting degenerate consensus sequence was $X^3X^{-2}C^{-1}$, which has little predictive power (Figure 3.1).

Experimentally, single nucleotide changes could not explain the influence of -3 to -1 sequences on reporter protein expression. For example, it might be suggested that an A at -3 enhances translation but a C at -3 does not, because ATC produced 388-fold higher reporter protein expression than CTC (Chapter Two, Figure 2.2). However, CCC or CTA at -3 to -1 resulted in high protein expression, indicating a C is not generally inhibitory at -3 (Chapter Two, Figure 2.2). We conclude that the effects of individual positions in the APR are context-
dependent. Since single nucleotide changes could not explain the influence of the APR, we examined di-nucleotide patterns in the APR using nearest neighbor analysis. Knowledge of nucleotide patterns that determine occupancy is important, because it can be used to make genome-wide predictions.

Popular tri-nucleotides (Figure 3.2) fell into the following patterns (Table 3.1): CXC, GXC, AXC, AXA, XTC, XCG, XCC, XAC, GCX, and ACX (Table 3.1). Tri-nucleotides absent from the data set belonged to the following rare patterns: TXT, TXA, CXT, AXT, XTT, XTG, XTA, XGT, XAT, TTX, TAX, CCX, and GGX (Table 3.1).

Di-nucleotide pattern usage depended on both the sequence and composition of the APR. For example, the frequently used motif CXC and the rare motif CCX have the same composition. Similarly, the popular pattern XTC differs from the rare pattern CXT by the order of the nucleotides at -2/-1 and -3 /-1, respectively.

The distributions of the expected and observed di-nucleotide frequencies for each set of patterns (-3/-1, -2/-1, and -3/-2) were compared by chi-square analysis (Figure 3.5). For each set, the difference in expected and observed di-nucleotide pattern frequencies was statistically highly significant (P < 0.01). We conclude that di-nucleotide pattern usage in the APR of *Leishmania* genes is not random.

**Biological significance of di-nucleotide motifs in APRs**

Our experimental work indicated that some popular tri-nucleotides (*e.g.*, ACC and GCC, Figure 3.2) resulted in robust translation in *Leishmania* (Chapter
Two, Figures 2.2 and 2.3). Incidentally, the genes for some highly expressed *Leishmania* proteins (e.g., gp63 and some tubulins) have GCC in their APR. From these data, we hypothesize that some tri-nucleotides are frequently used because they enhance protein synthesis. The rare tri-nucleotide CCC (Figure 3.2) resulted in high reporter protein expression, but GTA, another rare tri-nucleotide (Figure 3.2), produced low protein expression (Chapter Two, Figures 2.2 and 2.3). We hypothesize that some tri-nucleotides are absent from the data set because they result in extreme levels of translation, potentially acting as hyperactivators or very poor facilitators. These sequences might facilitate expression of abundant or scarce proteins as needed by the parasite.

These hypotheses can be applied to the di-nucleotide analysis. For example, ACC, GCC, and TCC are present from the frequently used XCC pattern, but CCC is absent. We hypothesize that ACC, GCC, and TCC result in high levels of protein synthesis, but CCC has an extreme effect on translation. The frequently used ATX pattern consists solely of ATC; ATA, ATG, and ATT are not observed in the data set. We hypothesize that ATC results in enhanced translation, but other ATX tri-nucleotides could be either very poor facilitators or hyperactivators. Similarly, some rare di-nucleotide patterns (CCX, TTX, XTA, and TAX) have only one member present in the data set. We hypothesize that these tri-nucleotides (CCG, TTA, GTA, and TAG) may result in high levels of protein synthesis while the other members of the patterns result in extreme levels of protein synthesis.
Experimental testing of the hypotheses will be a logical extension of this work. We will place a *Leishmania* 5’ UTR upstream of a reporter gene in a *Leishmania* expression plasmid containing a second reporter gene and a selectable marker gene (Chapter Four, Figure 4.1). Nucleotides in the APR of the 5’ UTR will be changed by site-directed mutagenesis. The effects of the sequences on reporter protein expression will be measured, and plasmid copy number per cell corrected by determining the amount of the second reporter protein. In this way, experimental investigation of the conclusions from genomic analysis presented herein may reveal the contribution of AUG-proximal regions to the regulation of gene expression in *Leishmania*. 
REFERENCES


Figure 3.1: Consensus APR from *Leishmania* 5' UTRs. The occurrence of each nucleotide (A, C, G, or T) at positions -3, -2, and -1 was analyzed for the 200 5' UTRs.
Figure 3.1

Occurrence in 200 5' UTRs

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td></td>
<td></td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>-2</td>
<td>20</td>
<td>15</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td>50</td>
</tr>
</tbody>
</table>
Figure 3.2: Frequency of tri-nucleotides in the APR of *Leishmania* 5' UTRs.

Sequences at -3 to -1 were obtained from the Sanger Centre *Leishmania major* genome project (http://www.sanger.ac.uk/Projects/L_major/). The 200 genes are found on chromosomes 1, 3, 4, 5, 19, and 23. 5' UTRs were selected only if the initiator AUG could be predicted by protein sequence similarity to a polypeptide from another organism.
Figure 3.2

-3 to -1 sequence

Occurrence in 200 *Leishmania* 5' UTRs
Figure 3.3: Comparison of expected and observed tri-nucleotide frequencies. The number of times a tri-nucleotide is expected in 200 independent trials was calculated using the binomial equation. The predicted distribution for success (x) between 0 and 17 was obtained. A) Expected and observed frequencies (P(x)) are shown. B) The expected frequencies were divided by the observed frequencies (P(x)) and the ratios were graphed.
Figure 3.3

A

Frequency of tri-nucleotides $P(x)$

- Observed
- Expected

Number of successes $(x)$

B

Observed/Expected tri-nucleotide frequencies $P(x)$

Number of successes $(x)$
Figure 3.4: Di-nucleotide patterns in the APR of *Leishmania* 5’ UTRs. The 64 possible tri-nucleotides were grouped into three sets of 16 di-nucleotide patterns. There are 48 total patterns consisting of A) di-nucleotides at -3 and -1, B) di-nucleotides at -2 and -1, and C) di-nucleotides at -3 and -2. For example, AAA, CAA, GAA, and TAA are possible members of the pattern XAA (where X is any nucleotide).
Figure 3.4

A

-3/1 Di-nucleotide patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXT</td>
<td></td>
</tr>
<tr>
<td>TXC</td>
<td></td>
</tr>
<tr>
<td>TXG</td>
<td></td>
</tr>
<tr>
<td>TAX</td>
<td></td>
</tr>
<tr>
<td>CXT</td>
<td></td>
</tr>
<tr>
<td>CXC</td>
<td></td>
</tr>
<tr>
<td>CXG</td>
<td></td>
</tr>
<tr>
<td>CXA</td>
<td></td>
</tr>
<tr>
<td>GXT</td>
<td></td>
</tr>
<tr>
<td>GXC</td>
<td></td>
</tr>
<tr>
<td>GXG</td>
<td></td>
</tr>
<tr>
<td>GXA</td>
<td></td>
</tr>
<tr>
<td>AXC</td>
<td></td>
</tr>
<tr>
<td>AXT</td>
<td></td>
</tr>
<tr>
<td>AXG</td>
<td></td>
</tr>
<tr>
<td>AXA</td>
<td></td>
</tr>
</tbody>
</table>

Occurrence in 200 *Leishmania* 5' UTRs

B

-2/1 Di-nucleotide patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXT</td>
<td></td>
</tr>
<tr>
<td>XTG</td>
<td></td>
</tr>
<tr>
<td>XTC</td>
<td></td>
</tr>
<tr>
<td>XTA</td>
<td></td>
</tr>
<tr>
<td>XCT</td>
<td></td>
</tr>
<tr>
<td>XCG</td>
<td></td>
</tr>
<tr>
<td>XCC</td>
<td></td>
</tr>
<tr>
<td>XCA</td>
<td></td>
</tr>
<tr>
<td>XGTT</td>
<td></td>
</tr>
<tr>
<td>XGG</td>
<td></td>
</tr>
<tr>
<td>XGC</td>
<td></td>
</tr>
<tr>
<td>XGA</td>
<td></td>
</tr>
<tr>
<td>XAT</td>
<td></td>
</tr>
<tr>
<td>XAG</td>
<td></td>
</tr>
<tr>
<td>XAC</td>
<td></td>
</tr>
<tr>
<td>XAA</td>
<td></td>
</tr>
</tbody>
</table>

Occurrence in 200 *Leishmania* 5' UTRs

C

-3/2 Di-nucleotide patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTX</td>
<td></td>
</tr>
<tr>
<td>TCX</td>
<td></td>
</tr>
<tr>
<td>TGX</td>
<td></td>
</tr>
<tr>
<td>TAX</td>
<td></td>
</tr>
<tr>
<td>CTX</td>
<td></td>
</tr>
<tr>
<td>CXX</td>
<td></td>
</tr>
<tr>
<td>CGX</td>
<td></td>
</tr>
<tr>
<td>CAX</td>
<td></td>
</tr>
<tr>
<td>GTX</td>
<td></td>
</tr>
<tr>
<td>GXX</td>
<td></td>
</tr>
<tr>
<td>GAX</td>
<td></td>
</tr>
<tr>
<td>ATX</td>
<td></td>
</tr>
<tr>
<td>AXX</td>
<td></td>
</tr>
<tr>
<td>AGX</td>
<td></td>
</tr>
<tr>
<td>AAX</td>
<td></td>
</tr>
</tbody>
</table>

Occurrence in 200 *Leishmania* 5' UTRs
Figure 3.5: Comparison of expected and observed di-nucleotide pattern frequencies. The binomial equation was used to calculate the number of times a di-nucleotide pattern is expected in 200 independent trials. For success (x) between 0 and 40, the predicted distribution was determined. Expected and observed frequencies (P(x)) for A) -3/-1 di-nucleotide patterns, B) -2/-1 di-nucleotide patterns, and C) -3/-2 di-nucleotide patterns are presented.
Figure 3.5

A

B

C
Table 3.1: Frequently used and rare di-nucleotide patterns. Frequently used patterns were found 20 or more times (at least half the maximum number observed) in the APR of 200 *Leishmania* 5' UTRs. Rare patterns (less than or equal to 1/10 the maximum number observed) were present four times or less.
<table>
<thead>
<tr>
<th>Positions</th>
<th>Frequently used di-nucleotide patterns</th>
<th>Rare di-nucleotide patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3 / -1</td>
<td>CXC, GXC, AXC, AXA</td>
<td>TXT, TXA, CXT, AXT</td>
</tr>
<tr>
<td>-2 / -1</td>
<td>XTC, XCG, XCC, XAC</td>
<td>XTT, XTG, XTA XGT, XAT</td>
</tr>
<tr>
<td>-3 / -2</td>
<td>GCX, ACX</td>
<td>TTX, TAX, CCX, GGX</td>
</tr>
<tr>
<td>Protein</td>
<td>Accession Number</td>
<td>-3 to -1</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------------------</td>
<td>-----------</td>
</tr>
<tr>
<td><strong>Chromosome 19</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Importin alpha-like protein</td>
<td>7630132</td>
<td>GCG</td>
</tr>
<tr>
<td>2 Possible CGI-81 protein</td>
<td>7630130</td>
<td>AAC</td>
</tr>
<tr>
<td>3 Translation release factor 1 homologue</td>
<td>7271236</td>
<td>ACA</td>
</tr>
<tr>
<td>4 Possible 7-TM transporter</td>
<td>7271235</td>
<td>GCC</td>
</tr>
<tr>
<td>5 Molybdopterin synthase sulphurylase</td>
<td>7271231</td>
<td>AGG</td>
</tr>
<tr>
<td>6 Probable inner dynein arm</td>
<td>7271226</td>
<td>ATC</td>
</tr>
<tr>
<td>7 Translation initiation factor 4E</td>
<td>7271225</td>
<td>AGC</td>
</tr>
<tr>
<td>8 Possible nucleosome assembly protein</td>
<td>7258888</td>
<td>GAT</td>
</tr>
<tr>
<td>9 Possible geranylgeranyl diphosphate synthase</td>
<td>7018579</td>
<td>CAC</td>
</tr>
<tr>
<td>10 ADP/ATP carrier, copy 2</td>
<td>7018578</td>
<td>AGG</td>
</tr>
<tr>
<td>11 Serine/threonine-protein kinase</td>
<td>7007335</td>
<td>GAG</td>
</tr>
<tr>
<td>12 Probable mitochondrial carrier protein</td>
<td>6580521</td>
<td>TCC</td>
</tr>
<tr>
<td>13 Possible 26S proteasome subunit</td>
<td>6580520</td>
<td>ATC</td>
</tr>
<tr>
<td>14 40S ribosomal protein S13</td>
<td>6635891</td>
<td>AAC</td>
</tr>
<tr>
<td>15 Possible synaptobrevin-type transport protein</td>
<td>6635890</td>
<td>ACC</td>
</tr>
<tr>
<td>16 Transcription factor IIS homologue</td>
<td>6635888</td>
<td>CAC</td>
</tr>
<tr>
<td>17 Lipoyl acid synthetase mitochondrial precursor</td>
<td>6635086</td>
<td>ACA</td>
</tr>
<tr>
<td>18 Probable kinase</td>
<td>6635085</td>
<td>CAG</td>
</tr>
<tr>
<td>19 Kinase</td>
<td>6635080</td>
<td>AGC</td>
</tr>
<tr>
<td>20 Possible pre-mRNA splicing factor</td>
<td>6635077</td>
<td>TCT</td>
</tr>
<tr>
<td>21 Possible peroxisome assembly factor-3</td>
<td>6635075</td>
<td>ACA</td>
</tr>
</tbody>
</table>

139
<table>
<thead>
<tr>
<th></th>
<th>Gene Name</th>
<th>Accession</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Probable protein kinase</td>
<td>5869899 GTC</td>
<td>TCGGACCCGCACCCCTCGCTGGTCGCGCGGACAGGGAAGCGCCGGGGTGCCTAG</td>
</tr>
<tr>
<td>23</td>
<td>Histone H3</td>
<td>5869895 AGT</td>
<td>CCCCCACCCCTACCTGCACTGGACAGCTCAGAGTCTCAGTATGG</td>
</tr>
<tr>
<td>24</td>
<td>Malate dehydrogenase</td>
<td>5852144 AAA</td>
<td>CACAGGTGATCATCCCTTCTCTGCGAACCTCCTCGGCCCTCAGAGAAA</td>
</tr>
<tr>
<td>25</td>
<td>Kinesin-like molecule</td>
<td>5852143 AGC</td>
<td>CTTTCCTGAGGCGACTGCGCTCCGAGATATACTGGAAGAGGACGCC</td>
</tr>
<tr>
<td>26</td>
<td>Probable glycerol 3 phosphate dehydrogenase</td>
<td>6572421 GAC</td>
<td>CTCGCTTCTCTGTGCTGATCGAGTTTCTCTCTGTGGGTTTTCTGGCGAC</td>
</tr>
<tr>
<td>27</td>
<td>Gamma tubulin related protein</td>
<td>6572417 GCC</td>
<td>CGCGCGTCTCCACAAAAAGAAAAGGACAGACACTGGGGGAAGAGGCG</td>
</tr>
<tr>
<td>28</td>
<td>Possible RNA helicase-like protein</td>
<td>662646 GTC</td>
<td>CACCCTCACACACACACACACACACACACACACACACACACGCAACACACTTGTCGTC</td>
</tr>
<tr>
<td>29</td>
<td>RNA polymerase subunit</td>
<td>662645 TCC</td>
<td>GCGAGCACCGACTCGACAAACACATTCAGGGTTATGAGTACCACACCTCC</td>
</tr>
<tr>
<td>30</td>
<td>Probable kinesin-like molecule</td>
<td>662643 AAC</td>
<td>GGACAGGGCCATCTGGCTACATCTGGCTACCTTGAAGCTGAAATGATGCACCAGAAA</td>
</tr>
<tr>
<td>31</td>
<td>Chromosome 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Elongation initiation factor 2, alpha subunit</td>
<td>7243854 TGG</td>
<td>AGGTTGGAGGCGAGACGAGGCTAGAATAACGACGCAGTCCAGAGTGG</td>
</tr>
<tr>
<td>33</td>
<td>Inorganic pyrophosphatase</td>
<td>7243847 CAG</td>
<td>AGGCATAGGCGTACCGGTCTTCCTCTGCGACTGAGCAGGACCTTACACGAG</td>
</tr>
<tr>
<td>34</td>
<td>RNA binding protein 1</td>
<td>7243846 AAA</td>
<td>GATCGCAAATTTCAACCGACAGCAAGCGCAACGAAAGGCGAAATACCCGCAAA</td>
</tr>
<tr>
<td>35</td>
<td>Leucine rich repeat protein 3</td>
<td>7243843 CGC</td>
<td>GCCCTCTCCTCTCCACCTCTCTCTCAGTGCCCTCCCCTCCTCCTCGGGC</td>
</tr>
<tr>
<td>36</td>
<td>6-phospho-fructo-2-kinase related protein</td>
<td>7143412 GTA</td>
<td>TACCTCTGTCTGTTGCGGCTGCGTTGTTTGTGTTGTTGAGCAGCTCCGTGA</td>
</tr>
<tr>
<td>37</td>
<td>Serine-threonine kinase 3</td>
<td>7143410 GCA</td>
<td>GTGTTGCTGTTGTCGTCGCCGCCCCTTCCTCCTCCTCCTCCTCCTCGAGCTGAGCA</td>
</tr>
<tr>
<td>38</td>
<td>26S protease regulatory subunit</td>
<td>7105525 GCG</td>
<td>CTCCATCGCGCCACCCCATCCCGCCGCCCTGCTGCGGCGACACACGCGG</td>
</tr>
<tr>
<td>39</td>
<td>U2 auxiliary factor large subunit</td>
<td>7105523 ACC</td>
<td>CACACCGCTCGCAAAGGCCAAAAGGAAATTCAGGTAGGAACGAGTACACCC</td>
</tr>
<tr>
<td>40</td>
<td>Phosphate-repressible phosphatase permease</td>
<td>7105521 ACG</td>
<td>TCTCCTCCTCCTCTCGCGCTCTCCTCTCGGGTACACACGCACTCCGAGCAAGGCG</td>
</tr>
<tr>
<td>41</td>
<td>Protein arginine methyltransferase</td>
<td>7025834 TCG</td>
<td>CGCAGAGGCATTGCTCCACCGGACTGTTGCTGCCCTGCTCGGTGACCTTCC</td>
</tr>
<tr>
<td>42</td>
<td>Helicase 1</td>
<td>7025833 CAG</td>
<td>CATCTCGCTGAGCCGTCGATCCGATCGCCGACCGGACCCGCGAAAGTCCGAG</td>
</tr>
<tr>
<td>43</td>
<td>Quinone oxidoreductase</td>
<td>7025831 GCC</td>
<td>GTCGACTGCGACATCTGAACCGGTGGCGACACAAACACACGAAAGGCC</td>
</tr>
<tr>
<td>44</td>
<td>Serine-threonine protein kinase 3</td>
<td>7025822 CCG</td>
<td>CGCGTCTACGGACACATGAGACGCGTACACCGCCACCTCAGCGCCCAGCGGCGG</td>
</tr>
<tr>
<td>45</td>
<td>Helicase 2</td>
<td>7025814 AAG</td>
<td>GCGCCGCGACGCCGACGAGGTCCGGTGAACACACGCGACACACGACGACAGGAA</td>
</tr>
<tr>
<td>46</td>
<td>Ankyrin repeat protein 1</td>
<td>7021714 GAC</td>
<td>AGATAGCAGAAGGCACAGAAGAAAGGCGTCGAGCCGTGCTGACGCTCCGTCGGTGAC</td>
</tr>
<tr>
<td></td>
<td>Protein Name</td>
<td>Accession</td>
<td>promoter</td>
</tr>
<tr>
<td>-----</td>
<td>--------------------------------------------------</td>
<td>------------</td>
<td>----------</td>
</tr>
<tr>
<td>46</td>
<td>Serine-threonine protein kinase 2</td>
<td>6978420</td>
<td>TGA</td>
</tr>
<tr>
<td>47</td>
<td>Ribosomal protein P1</td>
<td>6978413</td>
<td>GCC</td>
</tr>
<tr>
<td>48</td>
<td>Leishmania unknown transporter 2</td>
<td>6978410</td>
<td>CGA</td>
</tr>
<tr>
<td>49</td>
<td>Ribosomal protein L38</td>
<td>6899668</td>
<td>TCC</td>
</tr>
<tr>
<td>50</td>
<td>Serine-threonine protein kinase 4</td>
<td>6899658</td>
<td>ATC</td>
</tr>
<tr>
<td>51</td>
<td>Delta-1-pyroline-5-carboxylate dehydrogenase</td>
<td>6899657</td>
<td>AGG</td>
</tr>
<tr>
<td>52</td>
<td>U2AF 23/25 spliceosome component</td>
<td>6899656</td>
<td>GCC</td>
</tr>
<tr>
<td>53</td>
<td>ABC transporter protein 1</td>
<td>6778474</td>
<td>GCG</td>
</tr>
<tr>
<td>54</td>
<td>Leucine rich repeat protein 2</td>
<td>6778470</td>
<td>CGC</td>
</tr>
<tr>
<td>55</td>
<td>Cytochrome c oxidase assembly protein</td>
<td>6778468</td>
<td>TCA</td>
</tr>
<tr>
<td>56</td>
<td>DNA primase large subunit p58</td>
<td>6778467</td>
<td>ACC</td>
</tr>
<tr>
<td>57</td>
<td>Choline transporter protein 1</td>
<td>6778466</td>
<td>CAC</td>
</tr>
<tr>
<td>58</td>
<td>2-aminoethylphosphonate: pyruvate aminotransferase</td>
<td>6759104</td>
<td>TCG</td>
</tr>
<tr>
<td>59</td>
<td>Leucine Rich Repeat Protein 1</td>
<td>6630525</td>
<td>CTC</td>
</tr>
<tr>
<td></td>
<td><strong>Chromosome 4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Serine/threonine protein kinase</td>
<td>L779.1</td>
<td>GAC</td>
</tr>
<tr>
<td>61</td>
<td>Homologue of <em>C. elegans</em> ZK328.7 protein</td>
<td>L2969.1</td>
<td>GCC</td>
</tr>
<tr>
<td>62</td>
<td>Polypeptide deformylase</td>
<td>L4830.2</td>
<td>CGC</td>
</tr>
<tr>
<td>63</td>
<td>Homologue of <em>S. cerevisiae</em> YGR103W protein</td>
<td>L4830.3</td>
<td>ACA</td>
</tr>
<tr>
<td>64</td>
<td>Nascent polypeptide associated-complex protein</td>
<td>L4830.7</td>
<td>ACC</td>
</tr>
<tr>
<td>65</td>
<td>Nascent polypeptide associated-complex protein</td>
<td>L4830.8</td>
<td>TGC</td>
</tr>
<tr>
<td>66</td>
<td>60S ribosomal protein L10</td>
<td>L4830.9</td>
<td>AAC</td>
</tr>
<tr>
<td>67</td>
<td>E1-E2 type cation ATPase</td>
<td>L2385.6</td>
<td>AGG</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Accession</td>
<td>IUPAC</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------------------------------------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>68</td>
<td>Homologue of <em>S. cervisiae</em> sco1 protein</td>
<td>L2385.4</td>
<td>GAC</td>
</tr>
<tr>
<td>69</td>
<td>Fructose-1,6-bisphosphatase</td>
<td>L2385.1</td>
<td>AGC</td>
</tr>
<tr>
<td>70</td>
<td>tRNA nucleotidyltransferase</td>
<td>L1156.05</td>
<td>TCG</td>
</tr>
<tr>
<td></td>
<td><strong>Chromosome 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>Poly-A export protein</td>
<td>1407725</td>
<td>GCC</td>
</tr>
<tr>
<td>72</td>
<td>Ribosomal protein S7</td>
<td>2266903</td>
<td>ACG</td>
</tr>
<tr>
<td>73</td>
<td>Ribosomal protein S7</td>
<td>2266902</td>
<td>ACG</td>
</tr>
<tr>
<td>74</td>
<td>Pyrazinamidase/nicotinamidase</td>
<td>2266908</td>
<td>CGC</td>
</tr>
<tr>
<td>75</td>
<td>Long chain fatty acid acyl CoA synthetase</td>
<td>2266910</td>
<td>AAC</td>
</tr>
<tr>
<td>76</td>
<td>Long chain fatty acid acyl CoA synthetase</td>
<td>2266912</td>
<td>AAA</td>
</tr>
<tr>
<td>77</td>
<td>Long chain fatty acid acyl CoA synthetase</td>
<td>2266913</td>
<td>GCG</td>
</tr>
<tr>
<td>78</td>
<td>Long chain fatty acid acyl CoA synthetase</td>
<td>1617559</td>
<td>GCG</td>
</tr>
<tr>
<td>79</td>
<td>Long chain fatty acid acyl CoA synthetase</td>
<td>1617560</td>
<td>TCG</td>
</tr>
<tr>
<td>80</td>
<td>Mitochondrial tricarboxylate carrier</td>
<td>1617564</td>
<td>ACA</td>
</tr>
<tr>
<td>81</td>
<td>DNA damage inducible protein</td>
<td>2429118</td>
<td>GCA</td>
</tr>
<tr>
<td>82</td>
<td>150 KDa oxygen regulated protein</td>
<td>L3169.5</td>
<td>ACG</td>
</tr>
<tr>
<td>83</td>
<td>Mitochondrial processing protease</td>
<td>2266917</td>
<td>ACG</td>
</tr>
<tr>
<td>84</td>
<td>Ubiquitin activating enzyme</td>
<td>2266923</td>
<td>CCG</td>
</tr>
<tr>
<td>85</td>
<td>Serine/threonine protein kinase</td>
<td>3002481</td>
<td>TGC</td>
</tr>
<tr>
<td>86</td>
<td>Eukaryotic initiation factor 4A</td>
<td>3002484</td>
<td>ATC</td>
</tr>
<tr>
<td>87</td>
<td>Eukaryotic initiation factor 4A</td>
<td>3002483</td>
<td>ATC</td>
</tr>
<tr>
<td>88</td>
<td>Calcium activated potassium channel</td>
<td>3002487</td>
<td>ATC</td>
</tr>
<tr>
<td>ID</td>
<td>Gene Name</td>
<td>Accession</td>
<td>Exon 1</td>
</tr>
<tr>
<td>-----</td>
<td>-----------------------------------------------</td>
<td>-----------</td>
<td>-----------------</td>
</tr>
<tr>
<td>89</td>
<td>Calcium activated potassium channel</td>
<td>L9003.1</td>
<td>TCG</td>
</tr>
<tr>
<td>90</td>
<td>Peptidyl dipeptidease dcp, oligopeptidase A</td>
<td>L9003.2</td>
<td>AAC</td>
</tr>
<tr>
<td>91</td>
<td>Mitotic centromere-associated kinesin</td>
<td>2978452</td>
<td>CAG</td>
</tr>
<tr>
<td>92</td>
<td>Propionyl CoA carboxylase</td>
<td>2978458</td>
<td>ACC</td>
</tr>
<tr>
<td>93</td>
<td>Thymidylate kinase-related protein</td>
<td>2978458</td>
<td>GAT</td>
</tr>
<tr>
<td>94</td>
<td>Glutaredoxin-related protein</td>
<td>2978460</td>
<td>TCG</td>
</tr>
<tr>
<td>95</td>
<td>Chloride channel protein</td>
<td>2995581</td>
<td>TTC</td>
</tr>
<tr>
<td>96</td>
<td>Peptidyl-prolyl cis-trans isomerase, cyclophilin</td>
<td>2995581</td>
<td>ACC</td>
</tr>
<tr>
<td>97</td>
<td>DNA excision repair protein</td>
<td>2995587</td>
<td>ACC</td>
</tr>
<tr>
<td>98</td>
<td>Poly A export protein</td>
<td>1407725</td>
<td>GCC</td>
</tr>
<tr>
<td>99</td>
<td>Thioredoxin</td>
<td>2995590</td>
<td>GAG</td>
</tr>
<tr>
<td>100</td>
<td>Exopolyphosphatase</td>
<td>1407725</td>
<td>CAA</td>
</tr>
<tr>
<td>101</td>
<td>Putative ATP-dependent RNA helicase</td>
<td>CAC44739</td>
<td>GGC</td>
</tr>
<tr>
<td>102</td>
<td>Probable kinesin-like protein 5</td>
<td>CAC44743</td>
<td>TAG</td>
</tr>
<tr>
<td>103</td>
<td>Possible major ampullate spidroin 2-like protein</td>
<td>CAC44730</td>
<td>GCG</td>
</tr>
<tr>
<td>104</td>
<td>Possible RNA binding protein</td>
<td>CAC44733</td>
<td>AGA</td>
</tr>
<tr>
<td>105</td>
<td>Possible nitroreductase</td>
<td>CAC44734</td>
<td>ACA</td>
</tr>
<tr>
<td>106</td>
<td>Hypothetical ankyrin repeat protein</td>
<td>CAC44736</td>
<td>GCC</td>
</tr>
<tr>
<td>107</td>
<td>Possible kinesin-like protein</td>
<td>CAC44737</td>
<td>TAG</td>
</tr>
<tr>
<td>108</td>
<td>Dual specificity phosphatase</td>
<td>CAC32252</td>
<td>TCG</td>
</tr>
<tr>
<td>109</td>
<td>Possible GTP-binding protein</td>
<td>CAC32254</td>
<td>GCG</td>
</tr>
<tr>
<td>110</td>
<td>Possible viscerotropic leishmaniasis antigen</td>
<td>CAC32250</td>
<td>TTG</td>
</tr>
<tr>
<td>111</td>
<td>Hypothetical zinc f-CCCH protein</td>
<td>CAC32255</td>
<td>GCA</td>
</tr>
</tbody>
</table>

Chromosome 5
<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>Accession</th>
<th>Type</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>112</td>
<td>Probable lipoamide acyltransferase component</td>
<td>CAC32256</td>
<td>GTC</td>
<td>CCTGCCTGTGCGTTGCTCAAAGTCCGCAGGTCATCGCTTCCGGCCAGCGCTGTC</td>
</tr>
<tr>
<td>113</td>
<td>Possible paraflagellar rod protein</td>
<td>CAB89583</td>
<td>AGG</td>
<td>CACAAACCCCCACCACCATCGTGTCGTGCGTGTCGTCACATCAAAGGAGG</td>
</tr>
<tr>
<td>114</td>
<td>Possible non-canonical ubiquitin conjugating enzyme</td>
<td>CAB89584</td>
<td>CGC</td>
<td>TTGTGCTCCTCCTTCCCCCCCCCCACCCCAACCCCAACCCCATTTCTCCGCGC</td>
</tr>
<tr>
<td>115</td>
<td>Probable dipeptidyl-peptidase III</td>
<td>CAB89585</td>
<td>AAG</td>
<td>TAAACCAACCCCTCCGACAGATCCCGTGCTTTTGCTCTCTGACGCAAG</td>
</tr>
<tr>
<td>116</td>
<td>Possible 3-mercaptopyruvate sulphurtransferase</td>
<td>CAB89587</td>
<td>AGC</td>
<td>CCCTCTCCCTCCTCTCTCTAAACGCCCCCTCTTTCTCTCCGGTCCTCAGC</td>
</tr>
<tr>
<td>117</td>
<td>NADH dehydrogenase</td>
<td>CAC32247</td>
<td>CGA</td>
<td>CATTITGCCACTGCGCGATAGCCGCTGCAACTGCGGCGTCGACGCAACCGCGCA</td>
</tr>
<tr>
<td>118</td>
<td>Possible monocarboxylate transporter protein</td>
<td>CAC32260</td>
<td>GTC</td>
<td>TCACCAGGACAGCGGCTCCTCCCCCTCCTCCTCCCGACACCAAGGTCTG</td>
</tr>
<tr>
<td>119</td>
<td>MMR HSR1 GTP-binding protein</td>
<td>CAC32261</td>
<td>GAA</td>
<td>AGCCCTTTGAAATTTATTTATCTTTGACTTCTCACAGCGACCAAGGAGAA</td>
</tr>
<tr>
<td>120</td>
<td>Possible cytoplasmic dynein light chain</td>
<td>CAC32265</td>
<td>AGA</td>
<td>CTTTTCGGTGCCGGCTCCTCCGGCGCTTCTTTGTTGTTGTTGTTTTAGA</td>
</tr>
<tr>
<td>121</td>
<td>CDC27/NUC2-related protein</td>
<td>CAC32266</td>
<td>GTC</td>
<td>CCCTCCACCACCCACCCACCCCTCCCTCTCCCTCCGTCAGACCAACACGATC</td>
</tr>
<tr>
<td>122</td>
<td>ATPase alpha subunit</td>
<td>CAC32270</td>
<td>ATC</td>
<td>CCCCCCACCACCCACCCACCCCTCCCTCCTCCCGACCGACCACAGCATG</td>
</tr>
<tr>
<td>123</td>
<td>Hypothetical transmembrane protein</td>
<td>CAB89667</td>
<td>GTG</td>
<td>GGTGGTGTTGACATTTCTGTTGTTGGTTGCTATGCTTTCTCTGCCCCGCGG</td>
</tr>
<tr>
<td>124</td>
<td>Possible vacuolar ATP synthase subunit</td>
<td>CAB89670</td>
<td>AAA</td>
<td>GGCACCAGAGGTCGCGAAGGACCAACACCCAGACACGGCTGATTTCTGAAAAA</td>
</tr>
<tr>
<td>125</td>
<td>Possible long chain polyunsaturated fatty acid elongation enzyme</td>
<td>CAB89676</td>
<td>TGC</td>
<td>GCCCCGCGGTCTAGGCGCGGCGGCGTTGCGCTTACCTCCAGCTGATGTGTTGTC</td>
</tr>
<tr>
<td>126</td>
<td>Possible prefoldin-related protein</td>
<td>CAB89673</td>
<td>GCG</td>
<td>GGGTTACAGCGCGCTCCTTCAAATGAGCAAGAATAAAATAAGTAGCAGATAATC</td>
</tr>
<tr>
<td>127</td>
<td>Probable microtubule-associated protein homolog</td>
<td>CAB89601</td>
<td>GTC</td>
<td>CCAACAGACCCGACACCTCTTACGCGGCGTTGCAGCCCGACCGCTCGTC</td>
</tr>
<tr>
<td>128</td>
<td>Possible ATP-dependent RNA helicase</td>
<td>CAB89599</td>
<td>CGC</td>
<td>TGGTTCTGCTGCTCTCCTTACAGGTTCCCTCTGAGCACTACAGACGCGCGCC</td>
</tr>
<tr>
<td>129</td>
<td>Trypanothione reductase</td>
<td>CAB89598</td>
<td>CTG</td>
<td>CCCCCCTCCCTCCCCTCTCCGCAAGAAAATTTCTTCCGCTTCTCTCCGACGGCCTG</td>
</tr>
<tr>
<td>130</td>
<td>Possible ribosomal protein</td>
<td>CAB89597</td>
<td>TCG</td>
<td>ACCCCCTCTCCTCTTACGGCGCGCTCCTTTCTGTTATCCCTTCTTCCAAGTCTCAG</td>
</tr>
<tr>
<td>131</td>
<td>Possible minichromosome maintenance protein</td>
<td>CAB89596</td>
<td>TGG</td>
<td>CCTCCGGGCTGCGACGCGACGCTCAGCATCGCGGAAGACGATGCCTGTC</td>
</tr>
<tr>
<td></td>
<td>Possible glutaredoxin-like protein</td>
<td>CAB89595</td>
<td>TCC</td>
<td>CGACGTACGCCTGCTAGCCGGCCTGCTCATGCTGCTGCGCCGTCC</td>
</tr>
<tr>
<td>---</td>
<td>---------------------------------</td>
<td>-----------</td>
<td>-----</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>133</td>
<td>Possible mitochondrial phosphate carrier protein</td>
<td>CAB89593</td>
<td>CTC</td>
<td>TCCGTGACTCCGCTCCGCTCCGCTGGAGACCACCACCCTCCTGATTAATTCCTC</td>
</tr>
<tr>
<td>134</td>
<td>Possible tyrosine phosphatase-like protein</td>
<td>CAB89592</td>
<td>ACT</td>
<td>ACTCTCCCCCTTTAATCGCCCTGGAATTCAAGATACCGCTCACAAC</td>
</tr>
<tr>
<td>135</td>
<td>Possible cyclin 2-related protein</td>
<td>CAB88217</td>
<td>CTC</td>
<td>CTCCAGCAAGGGTGCTGTCGTCGTGTGCTGCCGACCCTCTCTCTC</td>
</tr>
<tr>
<td>136</td>
<td>Possible alpha collagen related protein</td>
<td>CAB88219</td>
<td>ACA</td>
<td>ACTTTTATCGAGCCCTACCTCCCCTCTCCCCCTCACCACACACACACACAC</td>
</tr>
<tr>
<td>137</td>
<td>Possible SRE-2 like transporter</td>
<td>CAB88220</td>
<td>AGC</td>
<td>GACGGTGCAGGAAAGCTTAAATGGTGCACAGCAGGCTCAGGC</td>
</tr>
<tr>
<td>138</td>
<td>5’-methylthioadenosine phosphorylase</td>
<td>CAB75631</td>
<td>GTC</td>
<td>CGCCACTCCCTTCACCCGACCCAGCCGAGGGCCG</td>
</tr>
<tr>
<td>139</td>
<td>Kinesin-like protein</td>
<td>CAB75639</td>
<td>CGG</td>
<td>CGGGGATCGTACGACTCTCAACAGAGCTGGAGAGGCCAGGGCCG</td>
</tr>
<tr>
<td>140</td>
<td>Epsilon-tubulin</td>
<td>CAC69155</td>
<td>GTC</td>
<td>CCTGCCGCTCGGTGCTGGACAAGCCGCTGGCGCTGGAGCCACCTGTCGTC</td>
</tr>
<tr>
<td>141</td>
<td>Alpha-tubulin</td>
<td>CAC69092</td>
<td>GCC</td>
<td>AACAAAAACCTTTAAGCCTCCTCTCCCTTCTCTCTCTTCAACCC</td>
</tr>
<tr>
<td>142</td>
<td>Beta-tubulin</td>
<td>CAB68715</td>
<td>ATC</td>
<td>CCTCCACACAAACACATACACCTTTTGTGCTGCGGCTCAGGCAG</td>
</tr>
<tr>
<td>143</td>
<td>Zeta-tubulin</td>
<td>CAB63128</td>
<td>GCC</td>
<td>GCCGGCGTCCGCAAAAGGGCAGAAACAGCTGGGAGAGCGCAG</td>
</tr>
<tr>
<td>144</td>
<td>gp63 pre-protein</td>
<td>CAA68673</td>
<td>GCC</td>
<td>CCCTGCGACAGCCAGAGCCCCCTCAGGAGGAGGGGCGG</td>
</tr>
<tr>
<td>Chromosome 23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>145</td>
<td>Possible mucin precursor</td>
<td>CAB71223.1</td>
<td>CAC</td>
<td>TGCAGCTGCTATGGAGCCGAGCAG</td>
</tr>
<tr>
<td>146</td>
<td>Possible RNA binding protein</td>
<td>CAB71228.1</td>
<td>ATG</td>
<td>GGCACCTTTTTGTTCTTTATCTGCCCACACGCTGGCGCTCC</td>
</tr>
<tr>
<td>147</td>
<td>Hasp A2</td>
<td>CAC44949</td>
<td>ACC</td>
<td>ACTACTCTTCCCTCTCTTATACAGGTTATATCTGTCCTCTTTACCC</td>
</tr>
<tr>
<td>148</td>
<td>Hypothetical TM protein</td>
<td>CAC44957.1</td>
<td>AGA</td>
<td>CACACATTAGTGAAAGGAAAGAGAAAAGACAGGAAAGTTATGGTACAG</td>
</tr>
<tr>
<td>149</td>
<td>Possible corinon</td>
<td>CAC44941</td>
<td>ATA</td>
<td>TTATACACTTTCTCTTCTAGATTGCTTGTCCTGCTCCAGTCAAC</td>
</tr>
<tr>
<td>150</td>
<td>Probable acid phosphatase</td>
<td>CAC44944</td>
<td>ACA</td>
<td>GTTTTCGCACTCCTCCGCTTACCTGGATGACGTGAGGTCCCAGGGTGTTCGAG</td>
</tr>
<tr>
<td>151</td>
<td>Possible methyltransferase</td>
<td>CAC44945</td>
<td>GGA</td>
<td>CCTGGCTGCTGCCGTTCCGCTCCTCTCCGCTCCAGGAGGGCAGAG</td>
</tr>
<tr>
<td>152</td>
<td>Probable t-complex protein</td>
<td>CAC44947</td>
<td>ACA</td>
<td>CTGCTGTGTGACTCTGACTCTGACTCTGACTCTGACTCTGACTCT</td>
</tr>
<tr>
<td>153</td>
<td>Possible surface antigen</td>
<td>CAB71273</td>
<td>GAC</td>
<td>TCTGTAACAAAGAAAGGTGACAGATGAGGGAAGCTTAACGAGAAATCAGG</td>
</tr>
<tr>
<td>154</td>
<td>Possible ATP DNA helicase</td>
<td>CAB71274</td>
<td>CGG</td>
<td>CATCCTGCTGGTCGCTGCGACCCTTCCTTGTGCTGCGCTGCGCTGAGG</td>
</tr>
<tr>
<td>155</td>
<td>PH domain protein</td>
<td>CAB71275</td>
<td>CGT</td>
<td>GCTTGGTTGTACGTTCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>156</td>
<td>Prefoldin subunit</td>
<td>CAB71282</td>
<td>AAA</td>
<td>AAAACACGCTGCTCCGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>157</td>
<td>WD40 repeat protein</td>
<td>CAB71230</td>
<td>GTC</td>
<td>TGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>158</td>
<td>Kinesin-related protein</td>
<td>CAB71236</td>
<td>GAG</td>
<td>TCGAGATGATGTCGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG</td>
</tr>
</tbody>
</table>

145
| 159 | Ubiquitin activating enzyme | CAB71237 | ATC | AGGTACCAGCAGACTCATACGAGCGGCGACCAAGGCGGGGGCATC |
| 160 | Possible actin interacting protein oxidase | CAB94667 | CGA | GCCCTCTCCCTTCCCCCTTGCCTGTTCCACAGGAGACGACGACGACGACGAG |
| 161 | Possible Na+/H+ antiporter | CAB94669 | GCG | GAGTCTTGCGACGCGACGCGGTGCGCCGCCATCGGCGGTGTGAGTCCG |
| 162 | Aldose 1-epimerase | CAB562836 | AAT | GTGCACTGCAAGGCGACGACGACGACGACCCGCTGCGTGTGAGTCCG |
| 163 | Possible crooked neck protein | CAB58379 | CAA | CATGAACTCAGTAAAAGAAGACCGAAAGCTGGGAGAGGGGCGTTAACCA |
| 164 | Probable farnesyl transferase | CAB58383 | GCG | TTGTCTTGTTACCGCGGCTTCCACGGAACCGGTCTTTGGGGCTGCG |
| 165 | Probable vacuolar ATPase proton pump | CAB58384 | ATC | TTCTTCTTTCTGTAGTCTGCTGTCATTTCCCGACTTGGCCATC |
| 166 | Possible proline synthetase associated protein | CAB58387 | GCG | GAATCAGCCATGGTGCAGACATCATCACAGAGGTCTCCGACGCGACTGCG |
| 167 | Serine hydroxymethyltransferase | CAB72302 | AGC | TCTGACGAGCTGAGCTTCCCCCTGAAAGTTGGCGACGCGGGCGATAGC |
| 168 | Translation initiation factor 3 | CAB72307 | ACC | CAAACAAACAGCAACCACACCGAAGGCGAGCTGCCATTTAGCCAGCCAG |
| 169 | DNA topoisomerase | CAB72310 | AAA | CAAGCATACTATACCGCAGAAGCAGCCAGCTGCTTTACACCAACGGGAAAA |
| 170 | Dynein heavy chain | CAB65927 | CGA | TCCCTCTTCCGCAACCCCTGGAAGCTAAAGATGGCGATGACAGAGGCGA |
| 171 | Possible exonuclease | CAB56931 | ATC | CTTCAGCTCCCTCCTCTCCTCTCCTGCTCTCGCCACATTTGCTGATAC |
| 172 | Possible cdc-2 related protein kinase | CAB59859 | GCC | CGTTTCGCAGCAGCAGCAGCAGCAGACCCGGTCGCGACCCGACCCGACCCGTCG |
| 173 | Possible drug resistance protein | CAB59836 | CAC | TCTTGTGGTAGGAACATCAGATCGTCTGTGAGACTGAGCACCACAC |
| 174 | Possible G10 protein homologue | CAB59864 | TGG | CTCCCTCCGCTCCCTTGGCGACACCTTGCCTCTGCTACCATGC |
| 175 | Possible MCM4 | CAB59865 | CAG | CATCTGACCTTCTCGAGATCTTCTCTATACCTACCCCAATCAAGCAGAC |
| 176 | Heat shock protein HSLU | CAB59867 | AAA | CGTGTCGGTGTCGACGCGAGCTTTTACTCTCCTTGCGAAAGGAAAA |
| 177 | Probable ATP-dependent permease precursor | CAB58407 | GTC | AAAGTCTCGCAGCAGGTAGCTGACAGACGACGTCTGCGGCTAACGTC |
| 178 | Possible MUS308 homologue | CAB58415 | AGA | TGACCGCCTTCTTCTCTCTATCCCGCAGGCTCGCTATACCTGACGAGC |
| 179 | Mitochondrial carrier protein | CAB58416 | GTG | CTGTTCAGCGCCGCGGAGAGCAGAAGACCCGACCCGACCCGACTTCCGTG |
| 180 | Probable DNA polymerase zeta catalytic component | CAB58420 | ACC | CTCCAGAAGCGAGCGCCTTATCGCTTCTCGCAGCAGACACC |
| 181 | Possible acetyltransferase | CAB85418 | CTC | TTTGCGCACTACATACACACGCTGCTGTCCTACCCCTTACCATGCTTG |
| 182 | Pteridine reductase 1 | CAB58390 | ACG | CCCCCTGTGGACCTGCTTAACTCGCAGGAGCAGTAATCGCTGACG |
| 183 | Tryptophanyl-tRNA synthetase | CAB58393 | TCT | CAGACACTCCCTCAAATCCCGCAGGCGCGCTCTGTGACTGCGTCTT |

146
| Probable G subunit of vacuolar type H+ ATPase | CAB58396 | ATC | AATCCCCCCCTCCCTCATCCCCCTGTTCCGAGTCGCGCTCATCATC |
| NADP-dependent alcohol hydrogenase | CAB58398 | GCC | CATCCCCCACCCTCTTTCCACTCACCACCAACGCTTTATCATCACCAGCC |
| Possible cytochrome C oxidase subunit 10 | CAB58399 | AGG | CTCTGATAAATCGAGCGCTATCCCCGGAAACGATAAAGGTATCATCAAGG |
| Possible acetyltransferase | CAB58288 | GCT | TCCGAGACGCGACGCGCGTAGTTAAAGGGAAGTAAACGTCATCCTTCGCT |
| Vacuolar type H+ ATPase | CAB58289 | ATC | CCGGCTTCCGCTGCTCCTGGTGCAGCTTCCGCTGCTCATCAGG |
| Possible cyclophilin | CAB58290 | ATC | CAGTACATCTTCGCATTCCGAAAAGTAAAGGTCGAGTCGAGTGTCGAGAGTCG |
| Possible lysosomal trafficking regulator | CAB58291 | TCG | TTGTCAAGCGCGGTCGTGGTCCTGCGCCACAAGAGCTGACGAGCGAGCGCTCG |
| Mannose-1-phosphate guanyltransferase | CAB58292 | AAG | TCGTCTCGTCTCCCCAGCAATACTCGCTTTCCATCTCAACCGCAAG |
| Possible arginase | CAB58296 | GCC | TTGTCCAGGGCAACACCAGCGTGACCGCCTAGGCCAGAGAGCTGCCC |
| Cyclophilin | CAB58298 | GGA | CGGTGGGGAAGAGGAGGAGGAGGATGGCGAGGGCTCATTCCGACGAGAG |
| Peroxidoxin precursor | CAB58299 | ACC | TTCTTCCGCTGCTACTTCTACCCGCGCGCGCGCTACCTCAAAGCAGCAC |
| Probable cytosolic aminopeptidase | CAB55612 | GCT | TCAGTCTCGAGGATCCGAGGATCCGATCCGTTTCGACACCCTCTAGACC |
| Possible brahma protein | CAB55616 | CCG | AGGATCGCGAGCGCGATCCCCATCTTGTCGAGTGAGTCAGTCGAGGCTTC |
| Probable sucrose-6-phosphate hydrolase | CAB55619 | ACG | TCCGCCCCCGACTCCTCGCTCTGTGGCGAGCGACAGCGACGAGGACG |
| Possible oxidoreductase | CAB55621 | TCC | AAACCACCACCCCACCCACAGAGCGATGCTGGCTTTAGTGAGGAAATCC |
| Probable T-complex protein 1 gamma subunit | CAB55542 | ACA | CTGCTCTCTGTACTTTACACCTCTCTCTTGAAAGGCGAAGCAATCAACA |
| Possible sirtuin | CAB55543 | ACG | GAAAATACGCGTTTCTTATTTATTATCGAGCTGTATCTCTTCCACG |
1 Conclusions

1.1 AUG-proximal regions of 5' UTRs affect translation in *Leishmania*

We are interested in understanding the effects of 5' UTRs on gene expression in *Leishmania*. The purpose of this study was to investigate the contribution of the AUG-proximal region (APR) (defined as the nucleotides at positions -3, -2, and -1 upstream of the translation initiation codon) to the regulation of *Leishmania* gene expression. In the context of a 180-nucleotide *Leishmania* 5' UTR [2], we tested eight 12-nucleotide sequences [1] differing only in the APR. The 12 nucleotide sequences consisted of (i) a polypurine hexamer (AGGAGG), (ii) an *E. coli* β-galactosidase lacZ spacer (CAG), and (iii) the APR [1]. Using stable transfectants of *Leishmania tropica*, we found that changing the sequence of the APR had remarkable effects on *Leishmania* protein expression. With one reporter gene, a 700-fold difference in reporter protein expression could be observed, and a second reporter gene yielded up to a 6000-fold difference. We identified three sequences, lacz-ACC, lacz-CCC and lacz-GCC, which enhanced translation of both reporter coding sequences. The reporter protein levels did not correlate with the amount of steady-state reporter mRNA. We concluded that the APRs influence gene expression post-transcriptionally, and explored the possibility that their action was at the level of translation. Polysome
analysis revealed that some APRs that enhanced protein expression were very efficient at recruiting ribosomes. One APR that inhibited protein expression acted after initiation of translation. From these data we concluded that APRs have the ability to influence protein synthesis in *Leishmania*. This is the first documented case of this mode of regulation in a trypanosomatid.

1.2 Sequences in the APR of *Leishmania* 5' UTRs are not random

The dramatic effects on translation from small changes in the AUG-proximal region encouraged us to study these sequences further. To that end, APRs from 200 *Leishmania* 5' UTRs were analyzed. Some tri-nucleotides (ACA, ACC, ATC, CAC, GCC, GCG, GTC) were over-represented in the data set, at frequencies up to $7.724 \times 10^5$ times higher than expected for random occurrence. Twenty-five percent of the possible tri-nucleotides were not found in the data set. The proportion of non-used tri-nucleotides (ATG, ATT, CAT, CCA, CCC, CCT, CTA, CTT, GGG, GGT, TAA, TAC, TAT, TGT, TTA, TTC, TTT) was 6 times higher than expected. The observed and expected distributions of tri-nucleotide frequencies differed significantly. We concluded that tri-nucleotides in the APR are not present by chance in the *Leishmania* genome.

Over-represented tri-nucleotides fit into ten frequently used patterns discovered by nearest neighbor analysis: AXA, AXC, CXC, GXC, XAC, XCC, XCG, XTC, ACX, and GCX (where X is any nucleotide). Absent tri-nucleotides belong to the rare patterns AXT, CXT, TXA, TXT, XAT, XGT, XTA, XTG, XTT, GGX, CCX, TAX, and TTX. The distributions of the observed and expected di-nucleotide pattern frequencies were significantly different. We hypothesize that
frequently used and rare di-nucleotide patterns in the APRs have physiological relevance in *Leishmania*.

1.3 Overall Conclusions

The experimental data presented in this study provides “proof of principle” that AUG-proximal regions can influence translation in *Leishmania*. The non-random occupancy of tri-nucleotides in the APR of *Leishmania* genes is consistent with our experimental data and suggests that the sequences may be biologically functional. In summary, this work highlights the potential of AUG-proximal regions to regulate gene expression in *Leishmania*.

2 Future Directions

2.1 Possible physiological relevance of tri-nucleotides at -3 to -1 of 5' UTRs

To further investigate the contribution of AUG-proximal regions (APRs) to translation in *Leishmania*, we will test hypotheses formulated from genomic analysis of -3 to -1 sequences. To facilitate the experimental work, a tricistronic plasmid, pNLG will be constructed (Figure 4.1). The plasmid will have two reporter genes, luciferase (*LUC*) and β-glucoronidase (*GUS*), and a selectable marker, neomycin phosphotransferase (*NEO*). A *gp63* 5' UTR will be placed upstream of *LUC*. Nucleotides in the APR of the *gp63* 5' UTR will be changed by site-directed mutagenesis. Stable transfectants will be used to study the effects of tri-nucleotides on *LUC* protein synthesis in *Leishmania*. The amount of GUSp will be used to control for plasmid copy number per cell.

Using pNLG we will test hypotheses outlined in Chapter Three (see Discussion) about the potential functional significance of tri-nucleotides in the
We hypothesize that over-represented tri-nucleotides lead to efficient protein synthesis in *Leishmania*, whereas absent tri-nucleotides result in extreme translation (acting as either hyperactivators or poor facilitators) (Figure 3.1). We will also test hypotheses derived from di-nucleotide analysis (Chapter Three, Figure 3.4 and Table 3.1). For example, from the frequently used ATX di-nucleotide pattern, only ATC is present in the data set, while ATA, ATG, and ATT are absent. We hypothesize that ATA, ATG, and ATT facilitate extreme levels of protein synthesis, and that ATC enhances translation in *Leishmania*.

### 2.2 Identification of cis-elements in *Leishmania* 5’ UTRs

Our long-term goal is to identify sequences in mature 5’ UTRs that contribute to the regulation of gene expression. To that end, we have examined the 200 *Leishmania* 5’ UTRs (Chapter Three, see Appendix) for putative cis-elements within -50 to -1 regions.

Comparison of the *Leishmania* -50 to -1 regions to eukaryotic cis-elements in UTRdb (http://bighost.area.ba.cnr/BIG/UTRHome/) produced only one hit. Several *Leishmania* 5’ UTRs have sequences related to the 15-lipoxygenase (LOX) differentiation control element (15-LOX DICE2) [3]. The sequence is located in the 3’ UTR of LOX mRNA, and facilitates tissue-specific translation of the enzyme. Binding of heterogeneous nuclear ribonucleoprotein (hnRNP) K and hnRNP E1 to 15-LOX DICE2 inhibits protein synthesis [4]. Fourteen *Leishmania* 5’ UTRs contain a DICE-related motif, *Leishmania*-DICE (*LeishDICE*), which contains a CU-rich region and an oligo-purine block with an AGXAGXXXG motif. We hypothesize that *LeishDICE* affects gene expression in the parasite. The
effects of *Leish*DICE (i) during differentiation and (ii) under various growth conditions will be tested using pNLG.

Other potential regulatory elements may be discovered by multiple sequence alignment of the -50 to -1 regions. Their effects on regulation could be discovered using the strategy proposed for *Leish*DICE. Studies of this nature are likely to make significant contribution to our understanding of the mechanisms of gene expression in *Leishmania*. 
REFERENCES


Figure 4.1: Tricistronic vector for gene expression studies in *Leishmania*. Plasmid pNLG is designed to have a portion of a *gp63* 5' UTR from *L. tropica* upstream of a luciferase (*LUC*) coding sequence. Dihydrofolate reductase/thymidylate synthetase (DHFR/TS) upstream (US) and downstream 1 and 2 (DS1 and DS2) sequences from *Leishmania* are required for pre-mRNA processing of neomycin (*NEO*), *LUC*, and β-glucoronidase (*GUS*).
APPENDIX A

CYSTEINE-LESS GPI-PHOSPHOLIPASE C IS INHIBITED COMPETITIVELY
BY A SULFHYDRYL REAGENT: GLYCO-MIMICRY BY PARA-
CHLOROMERCURIPHENYL SULFONATE

---

ABSTRACT

Glycosylphosphatidylinositol (GPI)-specific phospholipases are highly valuable for studying structure and function of GPs. GPI-phospholipase C (GPI-PLC) from Trypanosoma brucei and phosphatidylinositol-specific phospholipase C (PI-PLC) from Bacillus cereus are the most widely studied of this class of phospholipases C.

Sulfhydryl reagent inhibition of protein activity is indicative of cysteine (Cys) participation in biochemical events. The thiol reagent para-chloromercuriphenylsulfonate (pCMPS) inhibits T. brucei GPI-PLC which has eight cysteines. Surprisingly, we found that activity of B. cereus PI-PLC is blocked by pCMPS although the protein does not contain Cys. Inhibition of B. cereus PI-PLC was reversed when pCMPS was size-separated from a preformed pCMPS•PI-PLC complex. In contrast, no activity was recovered when T. brucei GPI-PLC was subjected to a similar protocol. Equimolar β-mercaptoethanol (β-ME) reversed inhibition of PI-PLC activity in a pCMPS•PI-PLC complex. For T. brucei GPI-PLC, however, ultrafiltration of the pCMPS•GI-PLC complex and addition of a large excess of β-ME was necessary for partial recovery of enzyme activity. Thus, T. brucei GPI-PLC is susceptible to inactivation by covalent modification with pCMPS, whereas PI-PLC is not. Kinetic analysis indicated that pCMPS was a competitive inhibitor of PI-PLC when a GPI was a substrate. Curiously, with PI as substrate inhibition was no longer competitive. These data suggest that pCMPS is a glyco-mimetic that occupies the glycan binding site of
PI-PLC from where, depending on the substrate, it inhibits catalysis allosterically or competitively.

**INTRODUCTION**

GPI-specific phospholipases have intriguing effects on cell physiology [1-4]. Phosphatidylinositol-specific phospholipase C (PI-PLC) from *Bacillus* spp. and GPI-PLC from *Trypanosoma brucei* are widely used for studying the biochemistry and cell biology of GPIs both *in vitro* and *in vivo* [5-8]. Consequently, understanding the biochemical and physiological roles of these enzymes is important. PI-PLC from *B. cereus* lacks cysteine (Cys) residues [9] whereas GPI-PLC from *T. brucei* has eight cysteines [10,11].

Sulfhydryl reagents are routinely used for determining whether (or not) cysteine residues are important for protein activity. For polypeptides that contain Cys residues it is generally thought that inactivation of their function arises from covalent modification of the polypeptide. While these concepts may be true in many instances other possibilities may be worthy of consideration.

First, inactivation of a protein by a sulfhydryl reagent may not necessarily mean that a Cys is essential for function. A case in point is GPI-PLC from *T. brucei*; the enzyme is inactivated by *para*-chloromercuriphenylsulfonate (*p*CMPS) [12,13], yet there is no evidence that a Cys is critical for catalysis [14]. Recently, Cys80 was identified as the putative target of *p*CMPS [15]. Cys80 is within a peptide stretch that has sequence similarity to PI-PLC from *Bacillus* species [16,17]. A Cys80Phe mutant of *T. brucei* GPI-PLC lacked enzyme activity, while
Cys80Ala mutant remains active but is highly resistant to \( p \)CMPS inactivation [15]. Coupled with the fact that Gln81 mutants of GPI-PLC were inactive, two factors were thought to account for the inactivation of GPI-PLC by \( p \)CMPS. First, proximity of Cys80 to the putative active site, and second, conversion of Cys80 to a bulky cysteinylmercuriphenylsulfonate adduct which could sterically hinder the active site [15].

Conceivably, sulfhydryl reagents could inhibit the function of cysteine-less proteins. An example of such a protein is the PI-PLC from \textit{Bacillus} spp. [18]. A logical interpretation of such data would be that the sulfhydryl reagent reacted with a non-Cys side-chain in the polypeptide.

We found that \textit{Bacillus cereus} PI-PLC is inhibited by \( p \)CMPS, as reported previously [18]. To explore how a sulfhydryl reagent inhibited a Cys-less protein, numerous hypotheses were examined. Several lines of evidence lead to a conclusion that \( p \)CMPS is a glyco-mimetic that competitively inhibits PI-PLC by binding to the glycan recognition site on the enzyme. In related studies, we established that \( p \)CMPS inactivated \textit{T. brucei} GPI-PLC by covalent modification.

**MATERIALS AND METHODS**

**Reagents**

\textit{para}-Chloromercuriphenylsulfonate (\( p \)CMPS), cupric sulfate (\( \text{CuSO}_4 \)), mercuric chloride (\( \text{HgCl}_2 \)) and [\( \text{I} \)-mercaptoethanol (\( \text{I} \)-ME) were obtained from Sigma Chemical company (St. Louis, MO, USA). Bovine serum albumin (BSA) was from New England BioLab (Beverly, MA, USA). Centricon-10 was purchased from
Amicon (Millipore, MA, USA) and [\(^3\)H]phosphatidylinositol (PI) was purchased from Amersham Pharmacia Biotech (Piscataway, NJ, USA). Soybean PI was obtained from Avanti Polar Lipids (Alabaster, AL, USA). EDTA disodium salt was from Fisher Scientific (Pittsburgh, PA, USA).

**GPI-phospholipase C Assays**

Recombinant *T. brucei* GPI-PLC from was purified from *E. coli* [19]. *B. cereus* PI-PLC (600 U/mg) was a gift from Dr. Mary Roberts (Boston University). Membrane form VSG (mfVSG) labeled with [\(^3\)H]myristate in its GPI (i.e., [\(^3\)H]mfVSG) devoid of SDS was isolated from *T. brucei* [9,9,19,20] and initially solubilized in 0.1% NP-40. It was diluted 40-fold in either a GPI-PLC assay buffer (50 mM Tris-HCl, pH 8.0, 5 mM EDTA, 1.0% NP-40) or in PI-PLC buffer (25 mM HEPES-KOH, pH 7.5, 0.1% sodium deoxycholate) before use. Enzyme assays were performed in 30 ml (GPI-PLC) or 40 ml (PI-PLC) with two mg of [\(^3\)H]mfVSG from 10 ml of appropriate assay buffer. Reactions were incubated at 37°C for 30 min unless otherwise stated.

**Preincubation of pCMPS with PI-PLC or GPI-PLC**

*B. cereus* PI-PLC (0.85 ng in 10 ml PI-PLC buffer) or *T. brucei* GPI-PLC (1 ng in 10 ml GPI-PLC buffer) was incubated with or without pCMPS (2 mM or 5 mM, respectively) in 20 ml at 37°C for 10 min. (pCMPS stock (100 mM) was prepared in 100 mM NaOH.)

**Removal of Unreacted pCMPS**

To prepare centricon-10 ultrafiltration units for use, sites for nonspecific protein adsorption were blocked with one ml of bovine serum albumin (BSA) (50 mg/ml
in water). After incubation of the phospholipase with pCMPS (in 20 ml), the mixture was diluted with 980 ml of the suitable buffer. The diluted mixture was centrifuged (5,000 x g, 4°C, 2.5 h) through a centricon-10 membrane. Approximately 50 µl of retentate was obtained for each enzyme, one-tenth of which was used per assay.

**Modification of [³H]mfVSG with pCMPS**

Twenty mg of [³H]mfVSG was incubated with (or without) pCMPS (2.5 mM final) in 20 ml of GPI-PLC buffer at 37°C for 10 min, after which 980 ml of chilled GPI-PLC buffer was added. [³H]mfVSG was recovered in the retentate after ultrafiltration through a centricon-10 membrane (5,000 x g, 4°C, 2.5 h). Two mg of pCMPS-treated [³H]mfVSG was used in specified GPI-cleavage assays.

**Inhibition of PI-PLC and GPI-PLC**

*B. cereus* PI-PLC (0.3 ng) was added to pCMPS (0.0625 mM to 0.625 mM) in PI-PLC buffer and kept at 37°C for 10 min. [³H]mfVSG was introduced, and the incubation continued at 37°C for 30 min. Released [³H]dimyristoylglycerol ([³H]DMG) was quantitated by liquid scintillation counting. For SDS inhibition of GPI-PLC, the detergent final volume (diluted from a 10% stock solution in assay buffer (see above)) was from 0.005% to 0.03%.

**[³H]PI substrate**

To prepare [³H]PI substrate (10 mM final concentration, 3000 dpm/µl), 54 µl of [³H]PI (55,500 dpm/µl) was added to 8.57 mg (9.99 µmol) of soybean PI in 100 µl chloroform [9]. The mixture was dried under nitrogen gas, and the PI resuspended in 1 ml of PI-PLC buffer. The suspension was sonicated for 5 min in
a Vibra Cell sonicator (Sonics & Materials, Pittsburgh, PA, USA) with amplitude at 20, and the pulser off. The substrate was stored at -20°C and thawed just before use.

[³H]PI digestion by PI-PLC and its inhibition by pCMPS, CuSO₄, and HgCl₂

_B. cereus_ PI-PLC (0.053 ng in 10 µl PI-PLC buffer, 0.15 nM final concentration) was incubated with pCMPS (0.5 µl to 20 µl of 10 mM, in PI-PLC buffer) at 37°C for 10 min [9]. (The amount of PI-PLC used was determined empirically to be within the linear region of the enzyme assay.) Ten µl [³H]PI (10,000 dpm) was added and the mixture (adjusted to a final volume 40 µl with PI-PLC buffer) incubated at 37°C for 30 min. The reaction was terminated with 250 µl of chloroform/methanol/HCl (66:33:1, v/v/v) [9,21]. Following centrifugation for 2 min (14,000 x g, 25°C), 75 µl of the upper phase containing [³H]myoinositol-phosphate was recovered for scintillation counting. For inhibition by heavy metals, PI-PLC was incubated with CuSO₄ (1.0-to-8.0 µM) or HgCl₂ (0.2-to-2.0 µM) in PI-PLC buffer, and reactants processed as described for the experiments with pCMPS.

**Reversal of PI-PLC inhibition by EDTA and β-ME**

EDTA (5-to-50 µM) or β-ME (10-to-100 µM) was added before incubation of PI-PLC with either pCMPS or HgCl₂. Following addition of [³H]mfVSG the mixture was incubated at 37°C for 30 min. The reaction (40 µl final volume) was terminated by addition of 500 µl water-saturated n-butanol, and centrifuged for 2 min (14,000 x g, 25°C). Four hundred µl of the upper phase was recovered for scintillation counting.
**Interfacial kinetics**

The scooting model of kinetic analysis (reviewed in [22]) was used because the substrates, \([^3]H\)mfVSG and \([^3]H\)PI, are cleaved at the interface between aqueous solution and detergent micelles. Effective inhibitor concentration at the micelle interface was expressed as a ratio of the mole fraction of inhibitor (\(X_i\)) to the mole fraction of the remainder of the components of the reaction mixture (1-X_i), including substrate, inhibitor, and detergent. Plots of \(V_o/V_i\) against \(X_i/1-X_i\) (\(V_o =\) reaction rate in the absence of inhibitor; \(V_i =\) reaction rate in the presence of inhibitor) were obtained. Competitive inhibitors are identified by two features: (1) data on the graph can be fit by a linear equation (with a coefficient of determination (\(r^2\)) of >0.90); (2) when \(X_i = 0\), \(V_o/V_i\) is close to one [22,23]. If either condition is not met, the points are fit with an exponential curve function of Cricket Graph III version 1.5.3 (Computer Associates International, Inc., Islandia, NY, USA), and the inhibition is characterized as not competitive.
RESULTS

pCMPS Inhibits a Cysteine-less PI-PLC From B. cereus

PI-PLC from B. cereus does not contain cysteine [18], yet the enzyme’s digestion of GPI is inhibited by the sulfhydryl reagent pCMPS (Figure A.1 A). The concentration of pCMPS required to inhibit PI-PLC by 50% (IC$_{50}$) was 0.3 mM, and 90% inhibition (IC$_{90}$) was obtained in 1 mM pCMPS.

Since $[^{3}\text{H}]$mfVSG contains Cys within the protein portion [24], inhibition of PI-PLC activity could arise from reaction of pCMPS with the substrate instead of PI-PLC. This possibility was checked with two approaches. In the first protocol, pCMPS-treated $[^{3}\text{H}]$mfVSG was used as substrate for the enzyme. pCMPS-treated $[^{3}\text{H}]$mfVSG was as good a substrate for PI-PLC as the unmodified $[^{3}\text{H}]$mfVSG (data not shown). This observation suggested that PI-PLC, not $[^{3}\text{H}]$mfVSG, was the target of pCMPS. In a second strategy, $[^{3}\text{H}]$PI, which does not contain Cys, was used as substrate. PI-PLC digestion of $[^{3}\text{H}]$PI was inhibited by pCMPS (Figure A.1 B). The IC$_{50}$ of pCMPS was 0.3 mM (Figure A.1 B), similar to the value obtained when the GPI-anchored $[^{3}\text{H}]$mfVSG was substrate (Figure A.1 A). The simplest interpretation of these data is that pCMPS interacts directly with PI-PLC during inhibition of enzyme activity.

Inhibition of a Cys-less PI-PLC by a thiol reagent was intriguing. Therefore, we investigated possible mechanisms by which pCMPS might act on the phospholipase C. We tested a hypothesis that inhibition of PI-PLC resulted from noncovalent interactions between the enzyme and pCMPS. For this, the protein was pre-exposed to the inhibitor, and residual pCMPS (0.415 kDa)
removed by ultrafiltration before assay of the “modified” enzyme obtained in a retentate (Figure A.2). In principle, enzyme activity would be inhibited if an adduct had formed between PI-PLC and pCMPS, whereas little loss of activity would be observed if no covalent adduct was created during the preincubation. Ultrafiltration of the PI-PLC•pCMPS complex resulted in loss of inhibition: PI-PLC activity was stimulated 10-fold (compare activity before and after filtration of pCMPS-treated PI-PLC (Figure A.2)). These data indicate that pCMPS inhibits PI-PLC without covalently modifying the enzyme.

**GPI-PLC from *T. brucei*: Removal of Unreacted *p*CMPS fails to Reverse Inhibition**

GPI-PLC from *T. brucei* contains eight cysteines [10,11], one of which appears to be at the enzyme active site [15]. The enzyme activity was inhibited completely by 5 mM *p*CMPS (Figure A.1 A). The IC$_{50}$ and IC$_{90}$ for *p*CMPS were 0.3 mM and 2 mM respectively.

We assessed whether inhibition of GPI-PLC correlated with covalent modification of the enzyme by *p*CMPS. When GPI-PLC was pretreated with *p*CMPS which was subsequently “removed” by ultrafiltration, no reactivation of enzyme activity occurred (Figure A.3 A). Hence, *p*CMPS is likely to have modified GPI-PLC covalently. Nevertheless, it was important to test whether modification of the substrate ([$^3$H]mfVSG) contributed significantly to loss of GPI cleavage activity in GPI-PLC. For this two substrates, *p*CMPS-treated [$^3$H]mfVSG and untreated [$^3$H]mfVSG, were used separately. *p*CMPS treatment slightly affected suitability of [$^3$H]mfVSG as a substrate, causing 33% loss of enzyme
activity (compare ‘untreated’ lanes in Figures A.3 A and A.3 B). This degree of inhibition is expected if the filtered [³H]mfVSG contained 13 mM unreacted pCMPS (Figure A.1 A; estimated from an initial 5 mM pCMPS, dilution before the filtration, and the final volume in the “retentate” (detailed in Materials and Methods)). We surmise that GPI-PLC, and not mfVSG, is the major target for pCMPS.

**B. cereus** PI-PLC: Equimolar [¶]-ME Reverses Inhibition by pCMPS

Further evidence for the non-covalent nature of the interaction between *B. cereus* PI-PLC and pCMPS was acquired by examining the effect of [¶]-ME on enzyme inhibition. ([¶]-ME can react with pCMPS.) PI-PLC was treated in two temporal stages, first with pCMPS (2 mM final), and then with [¶]-ME (2 mM). Substrate was added in the third stage, and GPI digestion determined.

Although introduced after exposure of the enzyme to pCMPS, [¶]-ME reversed the inhibition of PI-PLC by the sulfhydryl reagent (Figure A.4 A). (By itself [¶]-ME does not activate PI-PLC (Figure A.4 A).) To test whether [¶]-ME could protect PI-PLC from pCMPS inhibition, pCMPS was preincubated with [¶]-ME before addition of PI-PLC and substrate digestion. Equimolar [¶]-ME (*i.e.*, with respect to pCMPS) counteracted pCMPS completely (Figure A.4 B), but less than stoichiometric amounts failed to protect PI-PLC. This observation suggests that a “complex” formed between [¶]-ME and pCMPS cannot inhibit PI-PLC (Figure A.4 B).
**T. brucei** GPI-PLC: Excess [-ME is Required to Protect the Enzyme from pCMPS

The possibility that [-ME would reverse the inactivation of *T. brucei* GPI-PLC by pCMPS was investigated. GPI-PLC was treated with 5 mM pCMPS prior to addition of varying amounts of [-ME. ([-ME alone stimulated GPI-PLC two-to-threethree-fold (Figure A.5 A)). An equimolar amount (i.e., with respect to pCMPS) of [-ME failed to restore GPI-PLC activity (Figure A.5 A). A ten-fold excess of [-ME led to recovery of 50% of the maximal GPI-PLC activity (i.e., with [-ME). When pCMPS was preincubated with [-ME before addition to GPI-PLC, equimolar [-ME could not protect GPI-PLC from pCMPS inhibition (Figure A.5 B): only 13% of the maximal activity was recovered. A two-fold excess, at least, of [-ME was needed in preincubations with pCMPS to fully protect GPI-PLC (Figure A.5 B).

We tested a hypothesis that [-ME reversal of pCMPS inhibition was the result of reaction of the reducing agent with pCMPS-modified GPI-PLC. Towards this end, GPI-PLC was first treated with pCMPS (5 mM) and the complex filtered through a centricon-10 membrane. GPI-PLC in the retentate was treated with varying amounts of [-ME (Figure A.5 C). One mM [-ME restored 50% of GPI-PLC activity, and 5 mM was sufficient for 100% activity. In 50 mM [-ME, GPI-PLC activity was stimulated slightly (i.e., activity was above the basal level observed in absence of pCMPS) (Figure A.5 C).

When GPI-PLC was treated first with [-ME and the filtered enzyme studied, 1 mM of pCMPS inhibited the enzyme by 80% (data not presented). Thus, [-ME does not permanently alter the biochemical properties of
recombinant GPI-PLC: protection conferred by β-ME against pCMPS requires presence of the reducing agent.

Together, these data indicate that pCMPS, in contrast to its mode of action on B. cereus PI-PLC, modifies T. brucei GPI-PLC covalently. Consistent with this conclusion, a large excess of reducing agent is needed to restore partial enzyme activity to a pCMPS-GPIPLC adduct.

**pCMPS is a Competitive Inhibitor of B. cereus PI-PLC**

The absence of covalent modification during inhibition of PI-PLC by pCMPS made it possible to determine the nature of the inhibition by the thiol reagent. Interfacial kinetic analysis (reviewed in [22]) was chosen for this aspect of the work, since PI-PLC interacts with its substrates at the boundary between detergent micelles and the aqueous medium. Effective inhibitor concentration at the micelle interface was expressed as a ratio of the mole fraction of inhibitor (X_i) to the mole fraction of the remainder of the components of the reaction mixture (1 - X_i). Plots of V_o/V_i against X_i/1-X_i (V_o = reaction rate in the absence of inhibitor; V_i = reaction rate in the presence of inhibitor) [25] are presented (Figure A.6). Competitive inhibitors are identified by two features in the plots; (i) a successful fit of the data points to a linear equation (with a coefficient of determination (r^2) greater than 0.90); and (ii) When X_i/1-X_i = 0, the value of V_o/V_i approximates 1 (see Experimental Procedures). This approach cannot distinguish between non-competitive and uncompetitive inhibition.

pCMPS proved to be a competitive inhibitor of PI-PLC digestion of [^3H]mfVSG (Figure A.6 A). In control experiments, inhibition of T. brucei GPI-PLC
by SDS did not occur competitively; a linear curve fit to the data failed to satisfy
the required conditions (see earlier paragraph). An exponential curve fit for the
SDS inhibition is presented (Figure A.6 B). Surprisingly, when \[^{3}H\]PI was used as
substrate the linear equation obtained for the data was \( y = 6.105x - 1.831 \), and
the coefficient of correlation \((r^2)\) was \(0.886\) \((i.e.,\ less\ than\ 0.9)\). (That graph is not
presented.) When \(X_i/1-X_i = 0\), \(V_o/V_i = -1.831\) \((This\ value\ deviates\ significantly
from\ the\ expected\ value\ of\ 1.0)\). A much better fit of the data was obtained with
the exponential equation \(y = 0.893 \times 10^{0.559x},\ with\ a\ coefficient\ of\ correlation\ \(r^2 = 0.994)\) (Figure A.6 C). From the curve, \(V_o/V_i = 0.893\ \((i.e.,\ close\ to\ 1)\) \(When\ \(X_i/1-X_i = 0\). We conclude that when \[^{3}H\]PI is the substrate, \(p\)CMPS is not a
competitive inhibitor of PI-PLC.

**EDTA Distinguishes between \(p\)CMPS and \(Hg^{2+}\) Inhibition of \(B.\ cereus\ PI-PLC**

Mercury is a component of \(p\)CMPS. Consequently we considered the possibility
that inhibition of PI-PLC by the sulfhydryl reagent was due to \(Hg^{2+}\) contamination
of commercial \(p\)CMPS. (The manufacturer's product analysis sheet did not report
the presence of \(Hg^{2+}\)). We presumed that between 0.1-to-1% of Hg in \(p\)CMPS
was \(Hg^{2+}\) ion. Since 2 mM \(p\)CMPS inhibited \(B.\ cereus\ PI-PLC\ completely
(Figures A.1 A and A.1 B) we tested whether 2 mM \(Hg^{2+}\ \((i.e.,\ 0.1%\ of\ the
\(p\)CMPS concentration), presented as the chloride salt, would inhibit PI-PLC.
(Micromolar chloride ions, as the sodium salt, do not inhibit PI-PLC activity.)

Mercuric chloride was a potent inhibitor of PI-PLC; 0.6 mM inhibited PI-PLC by 50% (Figure A.7 A), as reported earlier [18]. In control experiments,
CuSO₄ inhibited PI-PLC feebly (Figure A.7 B). Although 5 mM reduced PI-PLC activity by 50%, susceptibility of the enzyme to Cu²⁺ tapered off; significant activity remained in 30 mM CuSO₄ (data not presented). Cu²⁺ is not as good an inhibitor of PI-PLC as Hg²⁺.

EDTA binds to metal ions. We therefore tested whether it could counteract the effects of Hg²⁺ on PI-PLC. When PI-PLC was preincubated with 10 mM Hg²⁺, equimolar EDTA blocked inhibition of PI-PLC (Figure A.7 C). This result is consistent with successful chelation Hg²⁺ by EDTA.

The effect of EDTA on ρCMPS inhibition of PI-PLC was different from that reported for Hg²⁺ (Figure A.7 C). Quantities of EDTA (5 mM and above) similar to those that rescued PI-PLC from Hg²⁺ inhibition (Figure A.7 C) failed to shield PI-PLC from ρCMPS (Figure A.7 D). If the ρCMPS effect on PI-PLC was due to contamination by 0.1% Hg²⁺ only 5 mM EDTA would reverse the inhibition. Yet 50 mM EDTA could not prevent ρCMPS inhibition of PI-PLC. These results indicate that the path of Hg²⁺ inhibition of PI-PLC is distinct from the mechanism employed by ρCMPS. More importantly, the data indicates that ρCMPS inhibition of PI-PLC is not the result of Hg²⁺ contamination of the sulphydryl reagent.

**DISCUSSION**

GPI-specific phospholipases are used routinely for discovery of GPI anchoring of proteins and polysaccharides to biological membranes [26], and as membrane-impermeable “probes” of GPI synthesis in vivo and in vitro [7,27-29]. The enzymes have been used to construct transgenic GPI-deficient cells to explore
the biological roles of GPIs [5,6,30]. These diverse uses of this unique class of phospholipases calls for clear understanding of the conditions that affect their activity in whole cells and in vitro.

Cysteine participation in biochemical and biological events is frequently revealed initially by susceptibility to sulfhydryl reagents (e.g., pCMPS or N-ethylmaleimide). Some recent cases involving membrane active proteins highlight interesting developments in the field. Calcium-insensitive cytosolic phospholipase A2 is inhibited by N-ethylmaleimide, yet mutagenesis studies indicate that none of the 9 Cys in the protein are critical for activity [31]. GPI-PLC from T. brucei is inhibited by sulfhydryl reagents [12,32,33], but no Cys is indispensable for enzyme activity [12,33,34]. Similarly, Cys residues in Na⁺/dicarboxylate co-transporter-1 are not needed for function although activity of the protein is blocked by p-chloromercuribenzenesulphonate [35]. At the other end of the spectrum a Cys-less PI-PLC can be inhibited by a sulfhydryl reagent (Figure A.1) [18].

Towards understanding how pCMPS affected B. cereus PI-PLC, our first objective was to determine whether the inhibition required covalent modification of the enzyme. In experiments where a preformed PI-PLC•pCMPS complex was resolved by ultrafiltration into enzyme (in a retentate) and free pCMPS (in the filtrate), most of the enzyme activity was recovered (Figures A.2 and A.4). This finding suggested that pCMPS inhibition of PI-PLC did not result from covalent modification. In a control experiment with T. brucei GPI-PLC, activity in a
pCMPS-treated enzyme could not be recovered in the retentate by following the protocols just described for PI-PLC (Figure A.3).

Further evidence of the absence of covalent modification during pCMPS inhibition of *B. cereus* PI-PLC activity was obtained from staged incubations of enzyme with either pCMPS or the reducing agent β-ME. After pretreatment of PI-PLC with pCMPS, the amount of β-ME needed to “reverse” inhibition was equimolar to the concentration of pCMPS (Figure A.4). One would not expect such low (relative) stoichiometry for the reducing agent if the enzyme had been modified covalently (see data on *T. brucei* GPI-PLC below). These data support the earlier conclusion (from the ultrafiltration experiments) that pCMPS does not modify *B. cereus* PI-PLC covalently.

With regard to the use of PI-PLC for analysis of GPIs, our observations suggest the following. First, it may be beneficial to avoid pCMPS in preparation of substrates from cell lysates. Second, inclusion of EDTA (50 mM) to PI-PLC buffers will be beneficial; EDTA is likely to keep the enzyme active in cases where small amounts of metal ions may have interfered with its action on GPIs. Third, if the use of sulfhydryl reagents is unavoidable in substrate preparation, investigators are encouraged to include β-ME in the reaction buffer to reverse inhibition by the thiol reagents.

Unlike *B. cereus* PI-PLC, *T. brucei* GPI-PLC displayed properties of a protein that was covalently modified with pCMPS. Following exposure to pCMPS, 5-to-10-fold excess of β-ME was needed to recover significant enzyme activity (Figure A.5 A). A preformed pCMPS•GPI-PLC complex had no activity even after
ultrafiltration to remove pCMPS (Figure A.5 C). Recovery of activity in the filtered enzyme (2.5 nM in the reaction) required a 2.5 x 10³ excess (1 mM) of β-ME (Figure A.5 C). Even then, 50 mM β-ME failed to restore full activity to GPI-PLC (not presented).

The nature of pCMPS inhibition of B. cereus PI-PLC (Figure A.6) is captivating. pCMPS has affinity for the GPI binding site of PI-PLC, since it is a competitive inhibitor of mfVSG digestion (Figure A.6 A). Curiously, with PI as a substrate the type of inhibition changes; it is no longer competitive (Figure A.6 B). The major difference between the two substrates (i.e., PI and GPI) is the presence of the EtN-phospho-glycan (EtN-Man₃-GlcN) group [27] on the GPI of mfVSG. So how does one reconcile the different modes of pCMPS inhibition?

PI-PLC binds to both the inositol and the glycan moieties of the substrate [36,37], and a GPI is a better substrate for PI-PLC than PI [38-40]. Scission of the phosphodiester bond at the inositol-1 position is influenced by glycan recognition. Surprisingly, data from crystal structures indicate that the myo-inositol head group of PI is not recognized within the same pocket where the glycan is bound to PI-PLC. myoInositol makes several contacts within a relatively deep groove in PI-PLC [37], but the glycan recognition site extends on shallow grooves along the surface of the protein (reviewed in [41]). From these considerations, an inhibitor that occupies the myo-inositol site need not interfere with another molecule binding to the glycan-binding site. Our data suggest that pCMPS binds the glycan recognition site on PI-PLC. Thus, when a GPI is the substrate the mode of inhibition is competitive because of the glycan on that substrate.
However, PI is recognized solely at the myoinositol site, meaning that PI does not compete with pCMPS which, acting as a glyco-mimetic compound, is at the glycan recognition site of PI-PLC. When PI is the substrate, pCMPS occupancy of the glycan binding site reduces rates of catalysis allosterically.

Our data on *B. cereus* PI-PLC and others like it [17] call for caution in interpretation of observations from sulfhydryl reagent inhibition of protein function. Where possible, supporting evidence for covalent modification (e.g. resuscitation with a reducing agent after removal of the inhibitor) can strengthen conclusions about how the sulfhydryl reagents act on the polypeptides. The experimental strategies employed in this study point to some rapid biochemical approaches for distinguishing between covalent and non-covalent inhibition of protein function by sulfhydryl reagents. Eventually, these biochemical studies can be complemented with mutagenesis of the Cys residues when genes for the proteins in question become available.
REFERENCES


Figure A.1: Inhibition of B. cereus PI-PLC and T. brucei GPI-PLC by pCMPS.

A) PI-PLC (0.85 ng) or GPI-PLC (1 ng) was preincubated with different concentrations of pCMPS in 20 µl of appropriate buffer (see Materials and Methods) for 10 min at 37°C. A 10 µl portion of [³H]mfVSG (2 µg, 18,500 dpm) was added and the mixture incubated for 10 min at 37°C. Released [³H]dimyristoylglycerol ([³H]DMG) was quantitated. Residual enzyme activity, as a fraction of the signal obtained without addition of pCMPS is presented. (One hundred percent activity corresponds to release of 9,000 dpm of [³H]DMG.) Data points are averages of duplicate determinations. All graphs presented in this manuscript are representative of three experiments (at least) that produced similar results. B) Bacillus cereus PI-PLC (0.053 ng) was incubated with pCMPS (0.1 mM-2.0 mM) in PI-PLC buffer at 37°C for 10 min. [³H]PI was added (100 nM final concentration, 10,000 dpm) and the reaction was kept at 37°C for 30 min. Released [³H]myoinositol phosphate was extracted with chloroform/methanol/HCl (66:33:1, v:v:v) and quantified by liquid scintillation counting. Residual (%) enzyme activity is presented. (100% digestion produced 4,500 dpm of [³H]myoinositol phosphate).
Figure A.1

(A) Residual (G)P-PLC Activity (%)

(B) Residual PI-PLC Activity (%)

$\rho$CMPS (mM)
Figure A.2: Effect of Ultrafiltration of a *B. cereus* PI-PLC•pCMPS Complex on Enzyme Activity. PI-PLC (24 nM) was treated with (or without) pCMPS (2 mM) prior to assaying for GPI cleavage. Where indicated, the mixture was filtered twice through a centicon-10 membrane (detailed in *Materials and Methods*). A portion of the retentate containing PI-PLC (2.4 nM final in the reaction) was assayed for GPI cleavage activity.
Figure A.2

![Bar chart showing PI-PLC Activity and [³H]DMG released (DPM).]

**PI-PLC:**
- pCMPS treatment: + - - +
- Filtered: - - + +
Figure A.3: pCMPS Inhibition of *T. brucei* GPI-PLC is Not Relieved by Ultrafiltration: *pCMPS*-treated VSG is a Good Substrate for GPI-PLC.  A)

Effect of Filtration of a GPI-PLC•pCMPS complex on Enzyme Activity. *T. brucei* GPI-PLC (25 nM) was preincubated with or without *pCMPS* (5 mM). The mixture was filtered through a centricon-10 membrane. A portion of the retentate containing GPI-PLC (2.5 nM) was used to cleave[^3]H]mfVSG, and released[^3]H]DMG quantitated.  B) *pCMPS*-treated[^3]H]mfVSG as a substrate for GPI-PLC.[^3]H]MfVSG was incubated with or without *pCMPS* (see Materials & Methods). The *pCMPS*•[^3]H]mfVSG mixture was filtered through centricon-10, and the retained[^3]H]mfVSG used as substrate for GPI-PLC. The enzyme itself had either been treated with *pCMPS* and filtered, or never exposed to the sulfhydryl reagent before filtration (see panel A).
Figure A.3

Experimental design (3B):

1. GPI-PLC + ρCMPS
2. 37°C 10 min
3. Ultrafiltration in centron-10
4. Add ρCMPS-treated [3H]mfVSG
5. 37°C 10 min
6. Quantitate [3H]DMG
Figure A.4: Equimolar [¶]-ME Counteracts pCMPS Inhibition of *B. cereus* PI-PLC.  

A) Preincubation of PI-PLC with *p*CMPS prior to [¶]-ME addition. PI-PLC (0.85 ng, 2.5 nM) was first treated with 2 mM *p*CMPS (37°C, 10 min). [¶]-ME (varying concentrations) was added (37°C, 10 min). [³H]mfVSG (2 µg) was introduced, and the mixture incubated at 37°C for 10 min. Released [³H]DMG was quantitated.  

B) Exposure of *p*CMPS to [¶]-ME before introduction of PI-PLC. *p*CMPS was pre-treated with different concentrations of [¶]-ME. PI-PLC was then introduced into the mixture. [³H]mfVSG was added finally and the solution incubated at 37°C for 10 min. Released [³H]DMG was quantitated as described in the legend to Figure A.1.
Figure A.4

**A**

Experimental design:

1. PI-P_C
2. Add β-ME
3. Add [3H]mVSG
4. Quantitate [3H]DMG

<table>
<thead>
<tr>
<th>Stage 1 (+ PI-PLC)</th>
<th>Stage 2 (+ β-ME)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 mM βCMPS</td>
<td>No βCMPS</td>
</tr>
<tr>
<td>10 mM β-ME</td>
<td>2 mM β-ME</td>
</tr>
<tr>
<td>5 mM β-ME</td>
<td>10 mM β-ME</td>
</tr>
<tr>
<td>2 mM β-ME</td>
<td>10 mM β-ME</td>
</tr>
</tbody>
</table>

**B**

Experimental design:

1. β-ME
2. Add βCMPS
3. Add PI-PLC
4. Add [3H]mVSG
5. Quantitate [3H]DMG

<table>
<thead>
<tr>
<th>Stage 1 (+ β-ME)</th>
<th>Stage 2 (+ PI-PLC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x buffer</td>
<td>No βCMPS</td>
</tr>
<tr>
<td>0.5 mM β-ME</td>
<td>2 mM βCMPS</td>
</tr>
<tr>
<td>1 mM β-ME</td>
<td>10 mM βCMPS</td>
</tr>
<tr>
<td>2 mM β-ME</td>
<td>0.5 mM β-ME</td>
</tr>
</tbody>
</table>
Figure A.5: *T. brucei* GPI-PLC: Effect of β-ME on pCMPS Inhibition.

A) Equimolar β-ME Fails to Reverse pCMPS Inhibition of *T. brucei* GPI-PLC. GPI-PLC (2.5 nM) was incubated with 5 mM pCMPS before addition of different concentrations of β-ME. [³H]mfVSG was added, and the reaction incubated at 37°C for 10 min. Released [³H]DMG was quantitated. B) Equimolar β-ME Partially Protects GPI-PLC from pCMPS Inhibition. pCMPS (5 mM) was first incubated with different concentrations of β-ME. GPI-PLC (2.5 nM) was added, followed by [³H]mfVSG whose digestion was carried out at 37°C for 10 min. [³H]DMG was quantitated after extraction into n-butanol. C) β-ME acts on pCMPS-modified GPI-PLC to Reactivate Enzyme Activity. *T. brucei* GPI-PLC (25 nM) was pre-incubated with or without 5 mM pCMPS at 37°C for 10 min. The mixture was filtered through centricon-10 (stage 1). GPI-PLC (25 nM) from the centricon-10 retentate was exposed to varying concentrations of β-ME at 37°C for 10 min (stage 2). [³H]mfVSG was added, and, the reaction incubated at 37°C for 10 min for GPI cleavage. Released [³H]DMG was quantitated.
Figure A.5

Experimental design:

A

GPI-PLC + βCMPS
37°C 10 min
Add β-ME
37°C 10 min
Add [H]mVSG
37°C 10 min
Quantitate [H]DMG

Stage 2
(+ β-ME)

Stage 1
(+ GPI-PLC)

5 mM βCMPS
No βCMPS

5 mM β-ME
25 mM β-ME
50 mM β-ME
No β-ME

GPI-PLC activity
[Hi]DMG released (DPM)

B

β-ME + βCMPS
37°C 10 min
Add GPI-PLC
37°C 10 min
Add [H]mVSG
37°C 10 min
Quantitate [H]DMG

Stage 2
(+ GPI-PLC)

Stage 1
(+ β-ME)

No βCMPS
5 mM βCMPS

5 mM β-ME
2.5 mM β-ME
5 mM β-ME
2.5 mM β-ME
10 mM β-ME
50 mM β-ME

GPI-PLC activity
[Hi]DMG released (DPM)

C

GPI-PLC + βCMPS
37°C 10 min
Ultrafiltration in centricon-10
Add β-ME
37°C 10 min
Add [H]mVSG
37°C 10 min
Quantitate [H]DMG

Stage 2
(+ GPI-PLC)

Stage 1
(+ GPI-PLC)

No βCMPS
5 mM βCMPS

5 mM β-ME
2.5 mM β-ME
5 mM β-ME
2.5 mM β-ME
10 mM β-ME
50 mM β-ME
Figure A.6: Nature of pCMPS Inhibition of *B. cereus* PI-PLC. (A and B)

[^3]H]mfVSG as Substrate. A) pCMPS inhibition[^3]H]mfVSG Digestion. Various concentrations of pCMPS were incubated with *B. cereus* PI-PLC at 37°C for 10 min. Following addition of[^3]H]mfVSG the mixture was incubated at 37°C 10 min. Released[^3]H]DMG was determined (see Figure A.1 A). The ratio of PI-PLC reaction rate in the absence of inhibitor (V₀) to the reaction rate in presence of inhibitor (Vᵢ) is plotted against the ratio of the mole fraction of inhibitor (Xᵢ) to the mole fraction of the remainder of the components of the reaction mixture (1-Xᵢ) (see Materials and Methods for details). B) SDS inhibition of GPI-PLCp. SDS was added to the reaction mixture during[^3]H]mfVSG digestion (see Materials and Methods), and the reaction analyzed as described in the legend to panel A. C)[^3]H]PI as Substrate. PI-PLC assays were performed and quantitated as described in Figure A.1 B. The ratio of the reaction rate in the absence of the inhibitor (V₀) to the reaction rate in the presence of the inhibitor (Vᵢ) was plotted against the ratio of the mole fraction of the inhibitor (Xᵢ) to the mole fraction of the remainder of the components of the reaction mixture (1-Xᵢ).
Figure A.6

A

$\frac{V_o}{V_i}$ vs. $\frac{X_i}{1-X_i}$

B

Inset diagram showing $V_o$ vs. $\frac{X_i}{1-X_i}$

C

$\frac{V_o}{V_i}$ vs. $\frac{X_i}{1-X_i}$
Figure A.7: Effect of EDTA on Hg$^{2+}$ and pCMPS Inhibition of *B. cereus* PI-PLC. 

A) HgCl$_2$ Inhibition of PI-PLC. PI-PLC was incubated with 0.2-to-2.0 μM HgCl$_2$ at 37°C for 10 min. [$^3$H]PI (100 nM, 10,000 dpm) was added, and the mixture incubated at 37°C for 30 min. Released [$^3$H]myo-inositol phosphate was quantified as described in the legend for Figure A.1. 

B) Effect of CuSO$_4$ on PI-PLC Activity. PI-PLC was incubated in 1.0-to-8.0 μM CuSO$_4$ at 37°C for 10 min. [$^3$H]PI was added and enzyme activity determined as described in A. 

C) EDTA Blocks HgCl$_2$ Inhibition of PI-PLC, but Cannot Prevent pCMPS Effect on PI-PLC. PI-PLC was incubated with 10 μM HgCl$_2$ or 5 mM pCMPS, and EDTA (5-to-50 μM) at 37°C for 10 min. [$^3$H]mfVSG (2 μg, 10,000 dpm) was added and the reaction was incubated at 37°C for 30 min. Released [$^3$H]DMG was extracted into n-butanol, and quantified by scintillation counting.
Figure A.7

[A graph showing the residual PI-PLC activity (%) against HgCl₂ (mM) with a secondary axis for CuSO₄ (mM) inset.]

Experimental design:
PI-PLC + ρCMPS or HgCl₂

± EDTA
37°C, 10 min
Add [³H]mVSG
37°C, 30 min
Quantitate [³H]DMG