The NADP-dependent glutamate dehydrogenase of *Giardia lamblia*: a study of function, gene structure, and expression

by

Janet Yee

B.Sc., The University of Toronto, 1983
M.Sc., The Queen's University, 1987

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in
FACULTY OF GRADUATE STUDIES
DEPARTMENT OF BIOCHEMISTRY
AND MOLECULAR BIOLOGY

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

July 1993

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of BIOCHEMISTRY AND MOLECULAR BIOLOGY

The University of British Columbia
Vancouver, Canada

Date JULY 29, 1993

DE-6 (2/88)

Abstract

Giardia lamblia is an interesting organism in several respects. Not only is it a medically important protozoan parasite, but its location in the phylogenetic tree is at a critical and informative position. Characterization of its subcellular structure and rRNA sequences suggest that *Giardia* is one the most primitive eukaryotes known. *Giardia* has an anaerobic metabolism that uses a NADP-dependent glutamate dehydrogenase (GDH) along with alanine aminotransferase to maintain an intracellular balance of $NAD(P)^+$-$NAD(P)H$ through the conversion of pyruvate to alanine.

In the initial part of this study, the *Giardia* NADP-GDH gene was cloned and characterized. The *Giardia* NADP-GDH is encoded by a single copy gene in the haploid genome. Transcript mapping and comparisons of the cDNA and genomic clones for the GDH gene did not detect the presence of introns or transsplicing. GDH transcripts have short untranslated sequences and are initiated only three to six nucleotides in front of the ATG translation initiation codon.

In the latter part of this study, the 5' flanking sequence of the *Giardia* GDH gene was analyzed to identify possible promoter elements since core promoter elements normally associated with RNA polymerase II transcription in higher eukaryotes are not found in the upstream sequence of the GDH gene or in other *Giardia* genes. Two novel sequence motifs were identified, an AT-rich element and a *Giardia* CAAT-box-like sequence called the g-CAB element. Their conservation and locations relative to the sites of transcription initiation suggest that these elements may be involved in the regulation of transcription. A 68 kD protein binds to the TTT trinucleotides found on either one or both strands of the DNA encoding these elements from the upstream region of the GDH gene. This 68 kD protein is referred to as POT for its poly(T) binding ability. The function of POT may be to help denature the DNA at the promoter region and to participate in the assembly of the RNA polymerase II pre-initiation complex in *Giardia*.

Table of contents

List of figures

List of tables

List of abbreviations

AMV	avian myeloblastosis virus
ATP	adenosine triphosphate
bp	base pair(s)
cDNA	DNA complementary to coding strand
CTP	cytidine triphosphate
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
dGTP	deoxyguanosine triphosphate
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
DTT	dithiothreitol
dTTP	deoxythymidine triphosphate
EDTA	ethylenediamine tetraacetic acid
GDH	glutamate dehydrogenase
HEPES	N-2-hydroxyethylpiperazine-N'-2-ethane sulfonic acid
kb	kilobase(s)
kD	kilodalton(s)
Mb	megabase(s)
MF3	muscle factor 3
MRE	metal response element
mRNA	messenger RNA
NAD(H)	nicotinamide adenine dinucleotide
NADP(H)	nicotinamide adenine dinucleotide phosphate
NTP	nucleoside triphosphate

PAGE	polyacrylamide gel electrophoresis
PBS	0.02 M potassium phosphate, pH 7.4; 0.155 M NaCl
PCR	polymerase chain reaction
PMSF	phenylmethyl sulfonyl fluoride
RACE	rapid amplification of cDNA ends
RNA	ribonucleic acid
RNase	ribonuclease
rRNA	ribosomal RNA
S	Svedberg unit of sedimentation coefficient
SDS	sodium dodecyl sulphate
Sp1	soluble protein 1
SSC (20X)	3.6 M NaCl; 0.3 M sodium citrate, pH 7
TAF	TBP associated factor
TBE	90 mM Tris-borate, pH 8.3; 2 mM EDTA
TBP	TATA-box binding protein
TE	10 mM Tris-HCl, pH 8.0; 1 mM EDTA
TEN	10 mM Tris-HCl, pH 8.0; 100 mM NaCl; 1 mM EDTA
TFII-I	RNA polymerase II transcription factor I
tris	tris (hydroxymethyl) aminomethane
tRNA	transfer RNA
USF	upstream stimulatory factor
V	volt(s)
VSP	variant-specific surface protein

Acknowledgments

I thank the members of my supervisory committee, Caroline Astell and Peter Candido for their help, advice and patience during the course of my studies. I thank my supervisor, Patrick Dennis, for giving me the opportunity to work in his lab, for introducing me to *Giardia* and S1 nuclease mapping.

I especially value the useful advice and protocols I have received from members of the Biochemistry Departments (Michel Roberge, Ivan Sadowski, Jan St. Armand, Nina Seto, Shu Hsu, Eve Stringham, Pat Tam, Don Jones, Steve, Rafferty, Louis Lefebvre) and other Departments (Rosie Redfield, Tom Cavalier-Smith, Rob McMaster, John Webb, Debbie Hays, Jennifer Couch).

Last, but not least, I thank the members of the Dennis lab (Peter Durovic, Phalgun Joshi, Willa Downing, Bruce May, Craig Newton, Deidre de Jong-Wong, Daiqing Liao, Shanthini Mylvaganam, Janet Chow, Luc Bissonnette, Josephine Yau) for their support and friendship. I would like to extend a special thank-you to Lawrence Shimmin whose natural curiosity and love of science was an inspiration.

<u>Dedication</u>

This thesis is dedicated to my parents, Ken and Vicky Yee, and to my husband, Steve Rafferty for their encouragement and support throughout my Ph.D. studies.

Introduction

Giardia lamblia is a protozoan parasite that infects mammals, reptiles and birds. *Giardia* infection is one of the most common causes of severe and chronic diarrhea in humans worldwide. In developing countries, infection of young children can lead to severe dehydration and malabsorption resulting in death. Despite its medical importance, information about this organism is limited due to the relatively recent development of an *in vitro* culture method for *Giardia lamblia* (Meyer, 1976).

Life cycle

During its life cycle, *Giardia* can exist in two forms, as a cyst or as a trophozoite. The cyst form is the infectious stage and the trophozoite form is the motile stage of the parasite. When a cyst is ingested by the host, by intake of either contaminated food or water, the cyst travels through the host's digestive system where it excysts in the small intestine. Two trophozoites emerge from one cyst after reproduction by binary fission and attach to the epithelium of the small intestine. During the course of infection, some of the trophozoites encyst and are passed into the host's stool. Infection of another host by these cysts allows the *Giardia* life cycle to be completed.

Cell structure

The *Giardia* cyst is egg shaped and its dimensions are approximately five by eight μm (Fig. 1C). As a cyst, *Giardia* is protected inside a cell wall that is highly resistant to environmental factors. The cell wall is composed of glycoproteins with galactosamine and glucose as the major sugar moieties (Sheffield and Bjorvatn, 1977; Jarroll *et al.*, 1989).
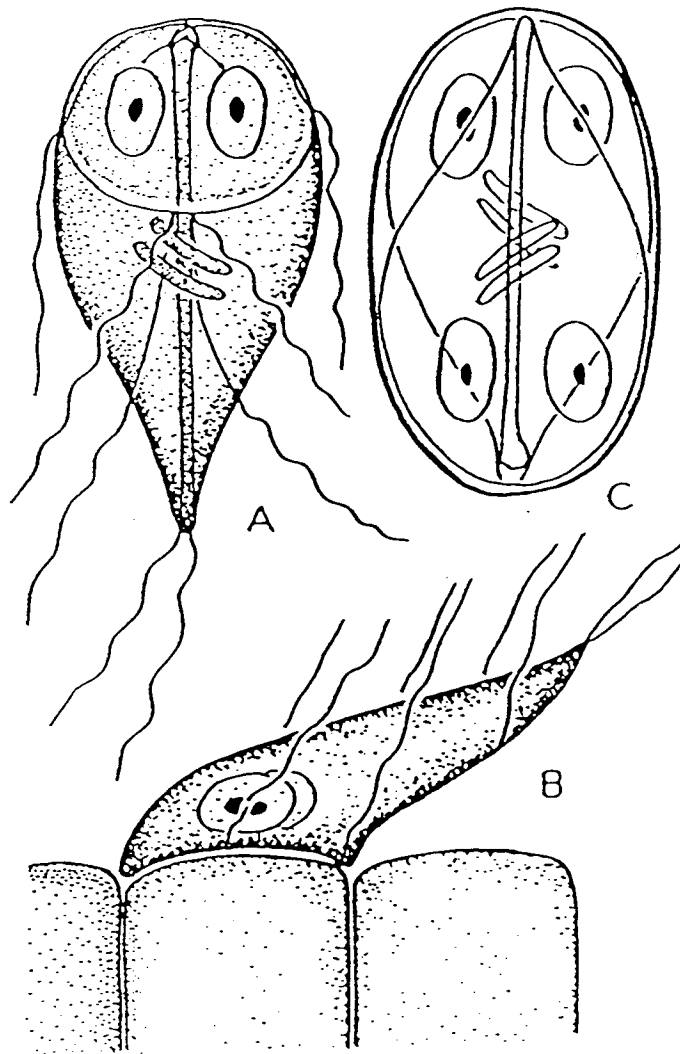
Figure 1: **Trophozoite and cyst forms of Giardia lamblia**
(A) The *Giardia* trophozoite. The ventral disk occupies the top third of the ventral surface of the organism. The two nuclei are shown underneath the ventral disk and the pair of median bodies are shown as rod-like structures in mid-body. (B) The *Giardia* trophozoite uses its ventral disk to attach to the surface of the epithelial cell. (C) The *Giardia* cyst. A *Giardia* trophozoite is replicated while protected inside the cyst's cell wall.

The *Giardia* trophozoite is tear-drop shaped, and its width and length are approximately five and ten µm, respectively (Fig. 1A). For its locomotion, *Giardia* makes use of four pairs of symmetrically arranged flagella. *Giardia*, like other eukaryotic cells, also has a nucleus, cytoskeleton, endoplasmic reticulum, and lysosomes, In fact, each *Giardia* has two nuclei which are identical in size, and are believed to be functionally equivalent (Kabnick and Peattie, 1990). A structure resembling a golgi apparatus is observed in *Giardia* during encystation (Reiner *et al.*, 1990).

Giardia has an unique disk-like structure on its ventral surface; this ventral disk is used as a suction device by the organism to attach to the host's small intestine (Fig. 1B). The contractile proteins, actin, myosin and tropomyosin, have been identified in the disk rim , and the structural proteins, tubulins and giardins, have been identified in the disk proper (Feely *et al.*, 1982; Peattie *et al.*, 1989). Giardins are proteins found exclusively in the *Giardia* ventral disks and range in size from 29 to 38 kD (Holberton, 1981; Crossley and Holberton, 1983; Crossley and Holberton, 1985; Holberton *et al.*, 1988; Peattie *et al.*, 1989). A structure referred to as a median body is also found only in *Giardia* trophozoites. This structure, found as a pair in the mid-body inside the organism, does not have a known function. However, the presence of giardins and tubulins in the median body suggest that it may be related to the ventral disk (Feely *et al.*, 1982; Crossley *et al.*, 1986).

The most interesting aspect of *Giardia* is the organelles it does not have. Although it is a eukaryote, it does not have mitochondria or peroxisomes. The lack of these organelles suggests that the *Giardia* lineage either lost them or diverged from the main eukaryotic lineage before these organelles were acquired by endosymbiosis (Cavalier-Smith, 1987).

Metabolism and glutamate dehydrogenases

Giardia is an aerotolerant anaerobe. Associated with its lack of mitochondria, *Giardia* lacks a Krebs' cycle and a cytochrome-mediated electron transport chain (Lindmark, 1980; Weinbach *et al.*, 1980). Most if not all of its energy requirements are met by anaerobic glycolysis (Muller, 1988). All the enzymes of carbohydrate metabolism in *Giardia* are present in the cytoplasm (Lindmark, 1988).

Glutamate dehydrogenases (GDHs) are enzymes that play important roles in carbohydrate and amino acid metabolism, as well as ammonia assimilation. These enzymes catalyze the interconversion between α-ketoglutarate and L-glutamate and utilize either NADP or NAD as coenzyme (Fig. 2). Based on their metabolic specificity and their oligomeric structures, GDHs can be classified into two types (Smith *et al.*, 1975).

Class I GDHs are hexameric proteins which are involved in ammonia assimilation; they are generally NADP-dependent but may also be NAD-dependent or even have dual coenzyme specificity. Class I GDHs are found in all organisms and the genes encoding these proteins are highly conserved between eubacteria and eukaryotes (Mattaj *et al.*, 1982; Hawkins *et al.*, 1989). This high level of sequence conservation, and the availability of homologous sequences from a number of different organisms representing all three kingdoms, makes class I GDHs good candidates for phylogenetic analysis (Benachenhou-Lahfa *et al.*, 1993).

Class II GDHs are tetrameric proteins which are NAD-dependent and are involved in the conversion of glutamate to α-ketoglutarate for catabolic purposes. The distribution of the class II GDHs are limited to some unicellular eukaryotes such as the yeasts, *Saccharomyces cerevisiae*, and *Neurospora crassa* (Vierula and Kapoor, 1989; Miller and Magasanik, 1990). All organisms characterized to date which have a class II GDH also have a class I GDH.
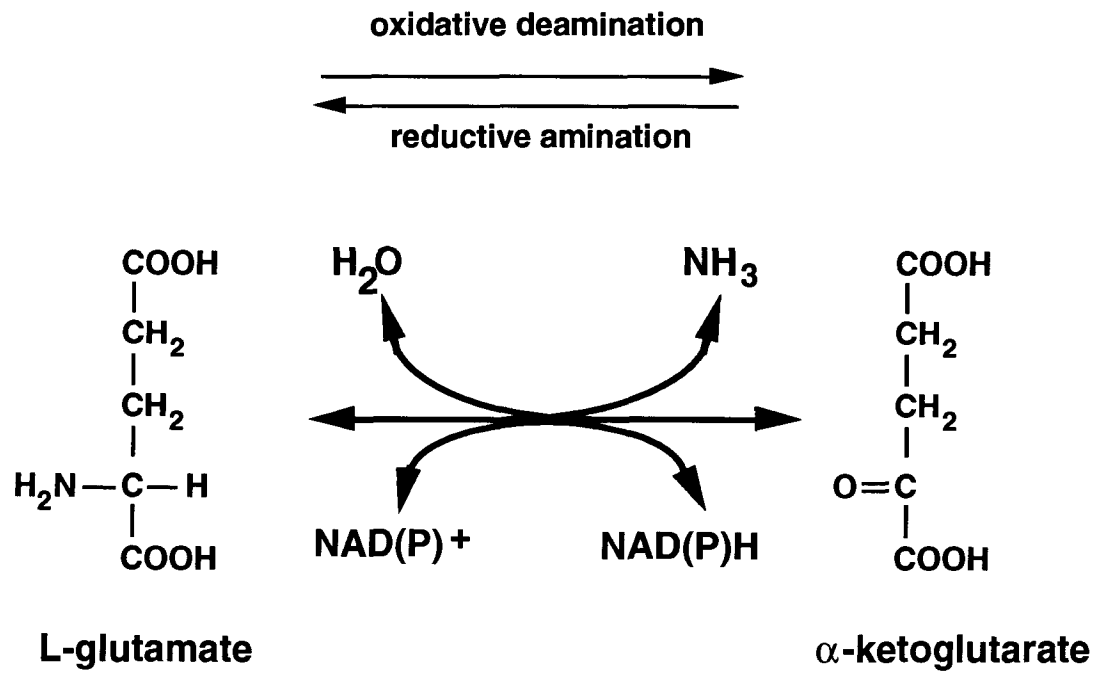
Figure 2: The reaction catalyzed by glutamate dehydrogenases
Glutamate dehydrogenase catalyzed the interconversion of L-glutamate and α-ketoglutarate using either NAD(H), NADP(H), or both as coenzymes.

Genome structure and organization

The number of different *Giardia* chromosomes has been estimated at four by microscopic examination of DAPI-stained nuclei and five by pulsed-field separation of chromosomes (Adam *et al.*, 1988; Kabnick and Peattie, 1990). Similarly, the genome size and the chromosome ploidy have not been resolved. Estimates of the haploid genome size range from as large as 80 Mb by C_0t analysis to as small as 12 Mb by size estimation of chromosomes separated by pulsed-field gel electrophoresis (Boothroyd *et al.*, 1987; Fan *et al.*, 1991). The estimation of chromosome ploidy has ranged from diploid to pentaploid (Adam *et al.*, 1988; Kabnick and Peattie, 1990).

The *Giardia* genes characterized to date have tended to be GC-rich; the G+C content is 67% for the β-tubulin genes and 75% for the rRNA genes (Kirk-Mason *et al.*, 1989; Healey *et al.*, 1990) Surprisingly, the overall G+C content of the *Giardia* genome was estimated to be only 48% by thermal denaturation analysis (Boothroyd *et al.*, 1987). These observations imply that the non-coding regions of the *Giardia* genome must be AT-rich if the interpretation of the melting profile is correct (Adam, 1991).

rRNA and evolutionary implications

The ribosome is a RNA-protein complex responsible for the translation of mRNA into proteins. Ribosomes can be separated into two components which are referred to as the large (L) and the small (S) subunits. The ribosomal RNAs (rRNAs) are named either according to their sedimentation coefficient (i.e. 5S, 16S, 23S), or according to their subunit association (i.e. the 16S rRNA is also referred to as the small subunit rRNA). The ribosomal proteins are named according to both their subunit association and their position relative to the other ribomsomal proteins after separation on a 2-dimensional polyacrylamide gel. Therefore, ribosomal protein L12 is associated with the large subunit and migrates to position twelve after 2-D gel electrophoresis.

Sequence analysis of its rRNA genes support the proposal that *Giardia* diverged early from other eukaryotes. *Giardia* has the shortest rRNA known for any eukaryote; the large and small subunit rRNAs are approximately 2300 and 1300 nucleotides long, respectively (Boothroyd *et al.*, 1987). These rRNAs are more similar in size to those from eubacteria and are missing regions which were once thought to be absolutely conserved among eukaryotes. Furthermore, sequence comparison of *Giardia* small subunit 16S-like rRNA to those from other organisms placed *Giardia* at the lowest branch of the eukaryotic phylogenetic tree (Fig. 3) (Sogin *et al.*, 1989).

Antigenic variation

The entire outer surface of the *Giardia* trophozoite, including the flagella, is covered by a variant specific surface protein (VSP). These proteins vary in size from 35 to >200 kD and contain 11 to 12% cysteine which are arranged in cys-X-X-cys motifs throughout the protein (Nash, 1992). *Giardia* possesses a family of genes encoding different VSPs but only one of these genes is expressed at any one time in an individual trophozoite. VSP expression changes spontaneously, and the frequency of change is a function both of the particular *Giardia* isolate and of the VSP (Nash *et al.*, 1990). This phenomenon is referred to as surface antigenic variation.

Although antigen variation is commonly thought to be a mechanism to evade immune response of the host, the biological significance of VSPs in *Giardia* has not been determined. *Giardia* isolates with different VSPs were observed to have variable susceptibility to intestinal proteases (Nash *et al.*, 1991). Recently, *Giardia* VSPs were shown to be able to bind zinc through their multiple cys-X-X-cys motifs (Nash and Mowatt, 1993). The biological importance of zinc binding is unknown but it was proposed that this may be a method used by *Giardia* to prevent the detrimental effects of oxidation (Nash and Mowatt, 1993). Zinc binding by VSPs has also been suggested to

0.1

Homo sapiens
Xenopus laevis
Saccharomyces cerevislae
Zea mays
Chlamydomonas reinhardtii
Ochromonas danica
Prorocentrum micans
Oxytricha nova
Plasmodium berghei
Dictyostelium discoideum
Euglena gracilis
Trypanosoma brucei
Vairimorpha necatrix
Giardia lamblia ◀━━━

Eukaryotes

Sulfolobus solfataricus
Halobacterium volcanii
Methanococcus vannielii

Archaebacteria

Bacillus subtilis
Escherichia coli
Anacystis nidulans

Eubacteria

Figure 3: **A phylogenetic tree derived from the analysis of the small subunit (16-18S) rRNA**

This unrooted phylogenetic tree contains organisms from the three domains: eubacteria, archaebacteria and eukaryota. The position of *Giardia* in the eukaryotic lineage is indicated by the arrow. The horizontal components (branches) of the tree represents the evolutionary distance between the organisms. The short bar indicated on the top left hand side of the figure corresponds to ten nucleotide changes per 100 positions. This figure presents the tree of Sogin *et al.*, (1989) derived by the distance matrix method.

be responsible for symptoms of zinc deficiencies in hosts infected with *Giardia* (Nash and Mowatt, 1993).

The control of antigenic variation in *Giardia* trophozoites is thought to be at the level of transcription or of RNA stability, since transcripts of expressed VSPs are observed whereas transcripts of non-expressed VSPs are not observed (Adam *et al.*, 1988; Adam *et al.*, 1992; Mowatt *et al.*, 1992; Nash and Mowatt, 1992). No other genetic alterations appear to be associated with expressed versus non-expressed VSP genes (reviewed by (Nash, 1992). At present, the mechanism controlling antigen variation has not been determined.

In the following sections, an overview of transcription initiation will be presented. Emphasis will be on specific aspects of this process relating to the analysis of putative promoter elements in *Giardia* genes which will be discussed in the second half of this thesis. These aspects include core promoter elements, transcription factors which bind these elements, and proteins which can bind single-stranded DNA.

Transcription initiation

Gene expression can be controlled at different levels by mechanisms that affect chromatin organization, DNA methylation, transcription, RNA decay, translation, post-translational modification, and protein degradation. The regulation of transcription is one of the early steps in the control of gene expression; the first step in the control of transcription is the initiation event, a multi-step process catalyzed by RNA polymerase. This process in eubacteria has been characterized most completely through studies on *Escherichia coli* (reviewed by Reznikoff *et al.*, 1985).

The complete *E. coli* RNA polymerase, called the holoenzyme, consists of five subunits ($\alpha_2\beta\beta'\sigma$) which can be separated into the core enzyme ($\alpha_2\beta\beta'$) and the sigma factor (σ). The core enzyme has an intrinsic ability to bind non-specifically to DNA and to initiate transcription randomly on the DNA template. The sigma factor is required to

help stabilize the binding of the enzyme to a specific DNA sequence and to begin transcription at a particular site, usually an A or a G nucleotide, in front of a gene. The region of the DNA which contains the sequences needed for the specific binding and transcription initiation by RNA polymerase is called the promoter.

The *E. coli* standard promoter contains two sequence motifs which are named according to their nucleotide position relative to the site of transcription initiation (Fig. 4A). The -10 sequence has the consensus, TATAAT, and the -35 sequence has the consensus, TTGACA. These consensus sequences are found in the majority of *E. coli* genes and the sigma factor that allows the RNA polymerase to recognize them is designated σ^{70} with the superscript indicating its molecular weight (in kD).

Transcription of some *E. coli* genes is induced under certain conditions and these genes have different consensus sequences in the -10 and -35 promoter region. Since there is only one type of core RNA polymerase in *E. coli*, different sigma factors are used by the RNA polymerase to recognize these alternative promoter elements. For example, the promoters of the genes induced by heat shock and nitrogen starvation are recognized by RNA polymerases with σ^{32} and σ^{60}, respectively (reviewed by Helmann and Chamberlin, 1988).

Eukaryotes, unlike eubacteria, have three different RNA polymerases to transcribe three different types of genes. Ribosomal RNA is transcribed by RNA polymerase I, mRNA by RNA polymerase II, and tRNAs and other small RNAs by RNA polymerase III. Similar to the eubacterial polymerase which requires a sigma factor, all three eukaryotic polymerases require accessory proteins for the initiation of transcription at the correct sites. These proteins are referred to as transcription factors. Formation of a transcription preinitiation complex requires the coordinate binding of these transcription factors to their respective recognition sites on the DNA template, and the recruitment of the correct RNA polymerase to this region (Buratowski *et al.*, 1989).

In RNA polymerase II transcription, transcription factors and sequence motifs that are found in the majority of genes and are sufficient to support a basal level of transcription are called general initiation factors and core promoter elements, respectively. Core promoters usually include two sequence elements called the TATA box and the initiator element, which are recognized by the transcription factors TBP and TFII-I, respectively (Fig. 4B).

In some genes, additional factors and their specific recognition sites are used to modulate transcription regulation. In mammals, the β-globin and the histone H2B genes have an additional promoter element, called the CAAT box, located between 40 to 100 bp upstream of the transcription initiation site (reviewed by La Thangue and Rigby, 1988). This element, with the consensus sequence GG(C/T)CAATCT, is bound by a family of transcription factors which up-regulate transcription of these genes.

TATA-binding protein (TBP)

The TATA-box sequence, with the consensus, TATA(A/T)AAT, is located 25 to 30 bp upstream of the transcription start site in the majority of RNA polymerase II promoters. The transcription factor which binds to this sequence is called TBP for TATA binding protein. TBPs are normally associated in a complex with a number of coactivators called TAFs for TBP associated factors (Nakatani et al., 1990; Peterson et al., 1990; Pugh and Tjian, 1990; Dynlacht et al., 1991). These TAFs-TBP associations are species specific and are indispensable for regulation by transcriptional activators (Pugh and Tjian, 1990).

The TBP gene has been cloned from a variety of eukaryotes including yeast, human, Drosophila, and the plant, Arabidopsis thaliana (reviewed by Dynlacht et al., 1991). Although the proteins encoded by these genes vary in size from 28 to 38 kD, they all share a highly conserved (>80% sequence identity over 180 amino acids) carboxy-terminal domain.

A: Eubacteria (E. coli) promoter

-35   -10   +1

TTGACA   TATAAT   A / G

B: Eukaryotic RNA polymerase II promoter

-25   +1

TATA A/T AAT   C/T A C/T TC C/T C/T C/T

TATA-box   **initiator element**
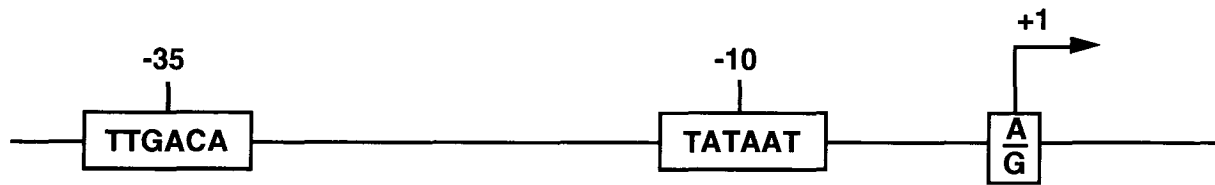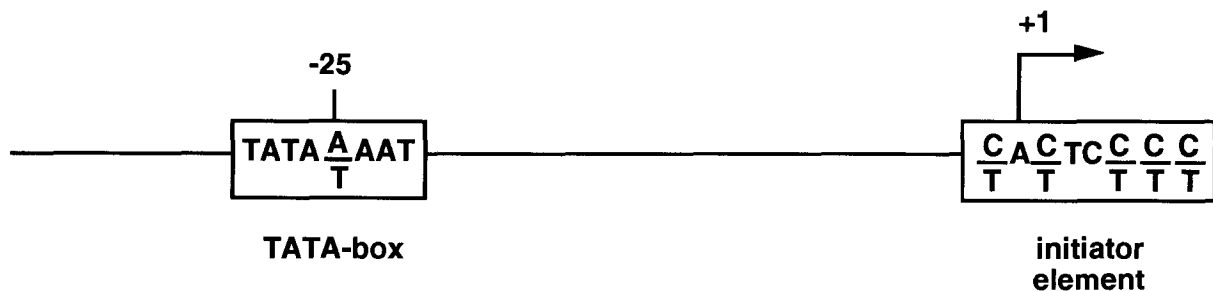
Figure 4: **A comparison of a eubacterial promoter and an eukaryotic RNA polymerase II promoter**
(**A**) An idealized eubacterial (*E. coli*) promoter with the -10 and -35 consensus sequences indicated relative to the site of transcription initiation, at a purine nucleotide, which is indicated by the arrow at +1. (**B**) An idealized eukaryotic RNA polymerase II promoter containing both a TATA-box and an initiator element. Transcription initiation, indicated by an arrow at +1, occurs within the initiator element.

This domain has been implicated in the DNA binding function of TBP (Horikoshi *et al.*, 1989; Hoffman *et al.*, 1990). The 180 amino acid carboxy-terminal domain of TBP contains a number of distinct motifs (Hoffman *et al.*, 1990). One of these is a set of amino acid repeats which contributes to the symmetry of the folded TBP protein. A second motif is a region with limited sequence similarity (23% identity and 40% similarity over 30 amino acids) to all bacterial sigma factors. This similarity is greatest in the alignment of the TBP sequence to that portion of the sigma sequence which is implicated in the binding of single-stranded DNA (Helmann and Chamberlin, 1988). A third motif consists of a repeat of basic residues which has the potential to form an α-helical structure with the basic residues located on one face. The fourth motif is a region that shows sequence similarity to the helix-loop-helix domains of a protein family related to the myc oncogene family of proteins.

In contrast to the high sequence conservation in the carboxy-terminus, the amino-terminal region of TBP from different eukaryotes varies markedly in both size and sequence and is responsible for the observed size variation of the protein. The amino-terminal of TBP is believed to be involved in interactions with transcriptional activators that are species specific (Pugh and Tjian, 1990).

It has long been known that TBP plays a pivotal role in the formation of the transcription preinitiation complex in TATA-box containing RNA polymerase II promoters, but recently, TBP was found to be required for RNA polymerase II promoters lacking a TATA-box, and also for RNA polymerase I and III promoters (Pugh and Tjian, 1991; Cormack and Struhl, 1992; Green, 1992; Sharp, 1992). How is TBP brought to the promoter region in the absence of a DNA binding site for this protein? While TBP binds directly to DNA at the TATA-box sequence in TATA-containing promoters, another transcription factor binds the DNA in TATA-less promoters and then recruits TBP directly or indirectly via yet another transcription factor to the

promoter region (reviewed by Rigby, 1993). For the RNA polymerase II TATA-less promoters, TFII-I and YY1 are two of these TBP tethering factors.

Initiator binding proteins and TBP-tethering factors

The initiator element, a pyrimidine-rich DNA sequence with the consensus, (C/T)A(C/T)TC(C/T)(C/T)(C/T), is located at the transcription start site in some TATA-containing and TATA-less RNA polymerase II promoters (Smale and Baltimore, 1989; Smale et al., 1990). A number of nuclear proteins from HeLa cells binds specifically to the initiator element. Two of these proteins are the transcription factors, TFII-I and YY1, which are approximately 120 and 68 kD in size, respectively (Roy et al., 1991; Shi et al., 1991).

In the TATA-containing adenovirus major late promoter, TFII-I was found to bind both the initiator and an upstream site called the E box, with the sequence CACGTG. Similarly, another transcription factor immunologically related to TFII-I, called USF (upsteam stimulatory factor), can also bind at both sites. TFII-I and USF were demonstrated to interact cooperatively at these sites by band-shift assays and by DNaseI footprinting. These results suggest that TFII-I binding at multiple promoter elements may facilitate interactions between upstream regulatory factors and the core promoter elements (Roy et al., 1991).

In the TATA-less promoter of the terminal deoxynucleotidyl transferase gene in mammals, the initiator has been implicated as being essential for core promoter strength and for determining the actual initiation site (Smale and Baltimore, 1989). These properties of the initiator were suggested to be a consequence of the recruitment of TBP to the promoter region by the binding of TFII-I to the initiator. (Roeder, 1991; Roy et al., 1991).

In several respects, YY1 is similar to TFII-I. Firstly, YY1 binds both at the initiator and at another upstream site in the adeno-associated virus P5 promoter (Shi et al., 1991).

Secondly, YY1 was demonstrated to be required for determining the transcription initiation site and promoter strength in both the adeno-associated virus P5 and adenovirus major late promoters (Seto *et al.*, 1991; Shi *et al.*, 1991). Lastly, YY1 can also bind at the initiator of the TATA-less terminal deoxynucleotidyl transferase gene promoter (Seto *et al.*, 1991). It is probable that YY1, like TFII-I, can also function as a TBP tethering factor in TATA-less promoters.

A subgroup of mammalian "housekeeping genes", which are expressed constitutively at low levels in all tissue types, have promoters that are TATA-less (reviewed by Sehgal *et al.*, 1988). These genes have another promoter element called the GC box, which has the consensus sequence CCGCCC. This element is usually found in multiple copies within 100 bp upstream of the transcription start sites. A transcription factor called Sp1 binds to the GC boxes and activates transcription (Dynan and Tjian, 1983; Dynan *et al.*, 1986). Another proposed function of Sp1 in these TATA-less promoters is to recruit TBP to this region indirectly by interacting with another transcription factor which has not yet been characterized (Pugh and Tjian, 1991). The identification of Sp1 has been restricted to mammalian cells although Sp1 can function as a transcription activator when introduced into a *Drosophila in vitro* transcription assay. In contrast, Sp1 does not affect transcription in yeasts in similar experiments (Pugh and Tjian, 1990). These observations suggest that Sp1 is unlikely to function as either a transcription activator or a TBP tethering factor in lower eukaryotes.

Protein binding to single-stranded DNA

Transcription initiation requires the unwinding of the duplex DNA in the promoter region to allow the RNA polymerase access to the single-stranded DNA template. Therefore, the ability to bind single-stranded DNA may be one of the properties of the RNA polymerase and any associated transcription factors (Horikoshi *et al.*, 1989).

The bacterial sigma factor has been shown by genetic criteria to bind single-stranded DNA (Helmann and Chamberlin, 1988). As the sigma factor allows the RNA polymerase to bind to the -10 and -35 sequences and to initiate transcription at the proper sites, the eukaryotic TBP recruits the RNA polymerase to the promoter region by binding to the TATA-box, allowing the formation of a preinitiation complex, and the start of transcription within the initiator element. Due to this functional analogy and its sequence similarity to the sigma factor, TBP may also be expected to bind single-stranded DNA (Horikoshi et al., 1989).

A protein called muscle factor three (MF3), which was isolated from an embryonic chicken muscle nuclear protein extract, can bind to single-stranded DNA containing promoter elements from muscle genes (Santoro et al., 1991). These promoter elements are the E-box motif from the mouse creatine kinase M gene, and the MCAT and MRE (metal response element) motifs from the chicken α-actin gene. The binding by MF3 to these elements was usually detected as a doublet in band-shift assays. In the separation of MF3 from other proteins in the nuclear extract by heparin-Sepharose chromatography, the activity that forms the lower band of the doublet elutes slightly before the activity that forms the upper band in band-shift analysis. Mutations in the MCAT motif which diminished binding by MF3 on band-shift assays also diminished the ability of the MCAT motif to activate transcription from a DNA construct containing the chloramphenicol acetyltransferase reporter gene in a transient transfection assay in muscle cells (Santoro et al., 1991).

Attempts to identify a common sequence required for MF3 binding to the three different promoter elements by interference footprinting and band-shift assays on mutant probes were unsuccessful. Despite the lack of sequence similarity among the different DNA probes recognized by MF3, the protein does discriminate in its binding to different DNA. MF3 binds both strands of the DNA containing the MCAT sequence, but binds only to the (-) strand of the DNA containing either the E-box or the MRE

sequence. Furthermore, MF3 does not bind to DNA containing other promoter elements such as the c-Fos and the cytoskeleton actin serum response element. Santoro *et al.* (1991) speculate that MF3 may recognize and stabilize altered conformations of DNA that may arise in the promoter region during muscle cell development.

MyoD is a transcription factor which can induce muscle differentiation in a wide variety of primary cells and transformed cell lines (Weintraub *et al.*, 1989). Similar to MF-3, MyoD can also bind single-stranded DNA such as the (-) strand of the DNA containing the creatine kinase M E-box. However, in contrast to MF-3, MyoD does not bind either DNA strand containing the MCAT or the MRE motifs (Santoro *et al.*, 1991).

Transcription initiation and promoters in *Giardia*

A limited number of genes have been characterized from *Giardia*; they include genes encoding giardins, tubulins, triosephosphate isomerase (TIM), ADP-ribosylation factor (ARF), and a number of variant cell surface proteins (Adam *et al.*, 1988; Kirk-Mason *et al.*, 1989; Gillin *et al.*, 1990; Mowatt *et al.*, 1992; Murtagh *et al.*, 1992). The characterization of the transcripts from these genes by a combination of S1 nuclease protection, primer extension and RNA sequencing indicate that the transcripts have a short 5' leader sequence that is initiated one to six nucleotides in front of the ATG translation initiation codon (reviewed by Adam, 1991). This is in contrast to higher eukaryotes where transcription initiates 20 to 100 or more nucleotides upstream of the translation initiation codon (Kozak, 1989). Furthermore, the examination of the regions upstream of the 5' transcript end sites has failed to identify core promoter elements associated with RNA polymerase II transcription in other eukaryotes.

Analyses of promoter sequences in the other protozoan parasites such as *Trypanosomes* and *Leishmania* have been confounded by the phenomenon of transsplicing, in which a leader sequence is added to the 5' end of every RNA transcript. Because this transsplicing process is thought to occur at the same time as transcription,

the identification of transcriptional start sites and associated promoter sequences has been unsuccessful in these parasites (Huang and Van der Ploeg, 1991). Since transsplicing has not been detected in *Giardia lamblia,* and sites of transcription initiation have been mapped for a number of genes, sequences associated with transcriptional regulation may be more easily identified and characterized in this organism.

Thesis Outline

This thesis has been organized into two parts, each with its own Results and Discussion sections. In the first part, I present the characterization of both genomic DNA and cDNA clones of the *Giardia* class I NADP-dependent GDH gene. Southern blot analysis and enzyme assays were performed to determine the copy number of the gene encoding the NADP-GDH and whether *Giardia* has a class II NAD-dependent GDH. Transcripts of the NADP-GDH gene were analyzed by primer extension and S1 nuclease protection experiments to determine the sites of transcription initiation and poly(A) addition. The protein sequence derived from the NADP-GDH gene was aligned to homologous proteins from *Neurospora crassa* and *E. coli.* to obtain an estimate of the degree of similarity among these proteins.

The second part describes my investigation of two conserved sequence motifs, referred to as the A/T-rich and the g-CAB elements, that I identified in the 5' flanking region of the GDH gene and other *Giardia* genes. The ability of proteins from a *Giardia* nuclear extract to bind specifically to DNA containing these elements in band-shift assays was investigated. The sequence specificity for protein binding was determined by cross-competition experiments and binding assays on DNA probes containing point mutations. The size of the protein bound to these DNA probes was determined by UV cross-linking experiments. These results and their significance will be discussed.

Materials and methods

Culture conditions

Giardia lamblia (Portland-1 strain, American Type Culture Collection) was grown at 37°C in modified TYI-S-33 medium (Keister, 1983). The composition of one litre of TYI-S-33 medium is as follows: 20 g casein digest, 10 g yeast extract, 10 g D-glucose, 2 g NaCl, 2 g L-cysteine HCl, 1 g K_2HPO_4, 0.6 g KH_2PO_4, 0.75 g bovine bile, 0.2 g L-ascorbic acid, 22.8 mg ferric ammonium citrate, and 100 ml heat-inactivated calf serum. All the above components, except for the serum, were dissolved in deionized H_2O and the pH of the solution adjusted to 7.0-7.2 with 4 M NaOH. The medium was then sterilized by passage through a 0.2 μm filter, and calf serum was added before use. To limit the amount of oxygen present in the culture flasks after inoculation, the flasks were filled to capacity with medium before capping.

Cells were harvested by chilling the culture flasks on ice for 15 to 30 min to allow detachment of the *Giardia* from the flask surfaces, followed by centrifugation of the culture at room temperature (1,200 rpm for 15 min in a Western H-103N centrifuge). The cells were washed twice in cold PBS buffer (0.02 M potassium phosphate, pH 7.4, 0.155 M NaCl).

Isolation of nucleic acids

For the isolation of DNA, washed *G. lamblia* cells were resuspended in TEN buffer (10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA) and lysed by addition of SDS (0.5%) and proteinase K (500 μg/ml). The mixture was incubated at 50°C for 1.5 h. The lysate was extracted once with phenol, once with phenol:chloroform (1:1) and once with chloroform. After dialysis overnight against TEN buffer at 4°C, the DNA was concentrated by ethanol precipitation and resuspended in TE buffer (10 mM Tris/HCl, pH 7.5, 1 mM EDTA). Contaminating RNA was removed by treatment with RNaseA.

For preparation of RNA, washed cells were lysed in 6 M guanidine HCl in 0.1 M sodium acetate at pH 5. The RNA was precipitated by the addition of 0.5 volume of ethanol and was centrifuged at 4°C for 1 h (14,000 rpm in a Beckman TL-100 ultracentrifuge).

<u>Amplification and cloning of PCR products</u>

PCR amplification was performed on 5 ng of poly(G) tailed cDNA using the RACE protocol as described by Frohman *et al.* (1988). The primers used were 5'GTCGAA(G/C)AG(G/C)CCGAA(G/C)CCCAT3' (oLW 72) and 5'GCGCTCTAGAC$_{14}$3' (adapter primer *Xba*I-dC$_{14}$). The PCR reaction was annealed at 57°C and extended at 72°C, repeated for 35 cycles. The amplification products were treated with DNA polymerase (Klenow fragment) and dNTPs, and ligated into the *Hinc*II site of the pUC 13 vector. One of the clones, designated pJYPCR10.5, was characterized further.

<u>Filter hybridizations</u>

All fragments that were used as probes in Southern filter hybridizations were purified from agarose or acrylamide gels and radiolabeled by the random primer method (Feinberg and Vogelstein, 1982). Unincorporated radioactivity was removed by either ethanol precipitation of the DNA, or by passage of the reaction mixture through a Sephacryl S-200 spin column (Pharmacia).

For Southern blot analysis, 8 µg of *Giardia* DNA was digested with various restriction enzymes and electrophoresed in a 0.8% agarose gel. The DNA was transferred onto Hybond-N membranes (Amersham) according to the manufacturer's instructions. In the analysis to determine the copy number of the NADP-GDH gene, the 2.4 kb *Pvu*II fragment containing the full length *Giardia* NADP-GDH gene was used to probe a genomic blot prepared as described above. After hybridization at 68°C, the blot was washed under conditions of high stringency (0.2X SSC, 0.1% SDS at 68°C). In the analysis to determine the presence of a *G. lamblia* gene homologous to the *Saccharomyces*

cerevisiae NAD-GDH, a 2.6 kb *Sal*I fragment containing about three quarters of the entire yeast genomic NAD-GDH gene was used to probe a *G. lamblia* genomic blot. The genomic clone containing the *S. cerevisiae* NAD-GDH gene was a gift from Steve Miller and Boris Magasanik (Miller and Magasanik, 1990). After hybridization at 42°C, the blot was washed initially at 2X SSC, 0.1% SDS at 42°C and subjected to autoradiography. The blot was rewashed with 2X SSC, 0.1% SDS at 52°C and autoradiography was repeated.

A *G. lamblia* genomic library in λEMBL3 and a *G. lamblia* cDNA library in λgt10 were gifts from D. Peattie (Peattie *et al.*, 1989). Duplicate filters from each library were probed at 60°C with ^{32}P-labeled cDNA insert from pJYPCR10.5. The filters from both libraries were washed at intermediate stringency (2X SSC, 0.1% SDS at 60°C); the genomic library filters were rewashed at higher stringency (0.2X SSC, 0.1% SDS at 65°C).

DNA sequencing

DNA fragments were cloned into pGEM-Zf vectors (Promega) and single-stranded templates were prepared for dideoxy chain termination sequencing (Sanger *et al.*, 1977) with T7 DNA polymerase. Regions of secondary structure were resolved by the incorporation of either deoxy-7-deazaguanosine triphosphates or deoxyinosine triphosphates in the sequencing reactions. Both strands of the genomic DNA and the cDNA containing the GDH gene were sequenced from overlapping clones obtained after progressive exonuclease III deletion of the full length DNA or cDNA clone (Henikoff, 1984).

Transcript mapping

To determine the location of the 5' transcript ends by primer extension, the oligonucleotide oJY1 (5'TCCACGACCTCCTCGACGGC3') was 5' end-labeled with T4 polynucleotide kinase and γ^{32}P-ATP. The radiolabeled oligonucleotide was hybridized

to 10 µg of *G. lamblia* total RNA and extended with AMV reverse transcriptase and dNTPs (Newman, 1987). The resulting extension products were electrophoresed on polyacrylamide gels alongside a sequence ladder generated using the same end-labeled oligonucleotide as a primer in a dideoxy chain termination sequencing reaction using an appropriate template.

Probes for S1 nuclease protection experiments were (i) a 441 bp *Hind*III-*Hinc*II fragment which was end-labeled on the (-) strand at the *Hinc*II site using T4 polynucleotide kinase and γ^{32}P-ATP and (ii) a 296 bp *Eco*RI-*Nar*I fragment which was end-labeled on the (-) strand at the *Eco*RI site by α^{32}P-dATP, α^{32}P-dTTP and DNA polymerase (Klenow fragment). The radiolabeled fragments were hybridized with 10 µg of *G. lamblia* total RNA and digested with S1 nuclease (Favaloro *et al.*, 1980). The products were electrophoresed on denaturing polyacrylamide gels alongside a molecular length standard (pGEM-7Zf(+), digested with *Msp*I and 3' end-labeled with dCTP and α^{32}P-dGTP), or a sequence ladder generated from the original radiolabeled fragment by the chemical degradation method performed on Amersham's Hybond M & G membranes (Maxam and Gilbert, 1980).

GDH sequence alignments

The protein alignment was performed using a DNASTAR alignment computer program utilizing a Lipman-Pearson algorithm. The alignment was checked by visual inspection.

To calculate the degree of amino acid identity in pairwise comparison, the sum of the number of positions with identical amino acids were calculated as a percentage of the total number of positions shared between the two sequences.

Enzyme assays

Washed cells were resuspended in 100 mM potassium phosphate buffer (pH 7.3) and lysed by two passages through a Yeda Press, once at 700 psi and once at 400 psi. The cell lysate was centrifuged at 4°C to remove particulate matter and cellular debris (12,000 rpm for 10 min in an Eppendorf 5412 centrifuge). For the forward reaction (the reductive amination of α-ketoglutarate) the supernatant was added to a glutamate dehydrogenase assay mixture containing 20 mM α-ketoglutarate, 20 mM NH_4Cl, and 0.1 mM of either NADPH or NADH in 100 mM potassium phosphate buffer, pH 7.3 . The assay conditions for determining endogenous background NADPH and NADH oxidation were the same as for the glutamate dehydrogenase assays except that 40 mM NaCl replaced the substrates, α-ketoglutarate and NH_4Cl.

For the reverse reaction (the oxidative deamination of L-glutamate), the supernatant of the *Giardia* lysate was added to an assay mixture containing 20 mM L-glutamate and 0.1 mM NAD^+ in 100 mM potassium phosphate buffer, pH 7.3. To determine endogenous background NAD^+ reduction, the same assay conditions were used except that 20 mM NaCl replaced the substrate L-glutamate.

The reactions were monitored at room temperature (on a Perkin-Elmer Lambda 3B spectrophotometer) by the change in optical density at 340 nm following the oxidation of NAD(P)H or the reduction of NAD^+. All assays were performed in duplicate. A unit of activity (U) is defined as the amount of enzyme which catalyzed the oxidation of 1 μmole of NAD(P)H or the reduction of 1 μmole of NAD^+ per minute under the assay conditions. Specific activity is defined as units per milligram of protein. The protein concentration of the crude extract was estimated by the Bradford assay using bovine serum albumin as the standard (Bradford, 1976).

Preparation of nuclear protein extracts

Nuclear extracts were prepared from freshly harvested *Giardia* cultures using a modification of the procedure of Andrews and Faller (1991). *Giardia* was lysed in hypotonic buffer (10 mM HEPES-KOH, pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 0.2 mM PMSF) and by the addition of 0.2% NP40. After incubation on ice for 10 min, a nuclear pellet was recovered by centrifugation of the cell lysate at 4°C (12,000 rpm for 15 sec. in an Eppendorf centrifuge). The nuclear pellet was resuspended in extraction buffer (20 mM HEPES-KOH, pH 7.9, 25% glycerol, 420 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 0.5 mM DTT, 0.2 mM PMSF), and incubated on ice for 20 min with occasional inversion of the sample tube. The extraction mixture was centrifuged for 4 min at 4°C (12,000 rpm in an Eppendorf centrifuge) and the extracted protein was recovered in the supernatant. The protein concentration of the extract was estimated by the Bradford assay using bovine serum albumin as the standard (Bradford, 1976). The extract was divided into small aliquots, frozen in a dry ice/ethanol bath, and stored at -70°C . The nuclear protein preparation procedure was monitored by fluorescent staining (Hoechst 33258 dye diluted 1:1000 from a 1 mg/ml stock solution) of the *Giardia* nuclei at various stages during its isolation and extraction.

Preparation of probes and competitors for band-shift assays

Table II lists the oligonucleotides used. Oligonucleotides were used either alone, as single-stranded DNA, or annealed together with the complementary strand, as double-stranded DNA. For annealing, the two complementary oligonucleotides were added in equimolar amounts to a 200 μl volume of an annealing buffer (50 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$, 50 mM NaCl), heated in a 90°C water bath for 5 min and cooled slowly to room temperature over 1 h. Both single-stranded and double-stranded oligonucleotides were radiolabeled with T4 polynucleotide kinase and $\gamma^{32}P$-ATP.

A 245 bp *Hind*III-*Sac*I fragment containing the GDH sequence (from -223 to +21 relative to the zero position assigned to A of the ATG translation initiation codon; or nucleotide position 1 to 245 in the sequence presented in figure 5) was cloned into a pGEM3Zf(-) vector (Promega). A 54 bp fragment, called GDH PCR54, was prepared from the pGEM clone by using oligonucleotides 2a and 1b as primers in a PCR amplification reaction. The primers were annealed at 50°C and extended at 72°C for 30 cycles in a 100 μl reaction. The 54 bp PCR product was gel purified and its sequence was checked by direct sequencing using the CircumVent thermal cycle dideoxy DNA sequencing kit (New England Biolabs). The fragment was radiolabeled by the substitution of half the total concentration of either dATP or dCTP with α^{32}P-dATP or α^{32}P-dCTP, respectively, in the PCR amplification reaction. For radiolabeling purposes, amplification was performed for 15 cycles in a 25 μl reaction. The radiolabeled GDH PCR54 fragment was gel purified prior to use.

Transcription of RNA *in vitro* was from the 245 bp *Hind*III/*Sac*I fragment cloned into the pGEM vector described above. The pGEM clone was linearized at the *Eco*RI site in the vector's polylinker and transcription was initiated at the SP6 promoter by the addition of NTPs and SP6 RNA polymerase. The RNA was radiolabeled by the incorporation of α^{32}P-CTP. The DNA template was removed after transcription by treatment with RQ1-RNase free DNase (Promega) and the RNA was gel purified prior to use.

Poly(dI-dC) and oligo(dT) 12-18 were obtained from Pharmacia.

<u>Band-shift assays</u>

Binding reactions contained 2.5 μg of poly(dI-dC), 3-5 μg of nuclear extract protein, and 0.02-0.05 pmoles of radiolabeled probe in a 20 μl volume of binding buffer (20 mM Tris-HCl, pH 7.5, 100 mM NaCl, 10% glycerol, 1 mM DTT, 0.2 mM PMSF). In competition experiments, all components, including probe and competitor DNA, were

pre-mixed before the final addition of the protein extract. In the studies on single-stranded DNA binding, the poly(dI-dC) was first heated at 100°C and cooled quickly on ice prior to its addition to the binding reaction. The reactions were incubated on ice for 20 min. A 5% polyacrylamide gel was pre-electrophoresed for 1-2 h at 200 V in 0.5X TBE/1% glycerol with a replacement of fresh running buffer prior to sample loading. Samples were loaded onto the gel and subjected to electrophoresis at 150 V for 1.5 - 2 h at room temperature. After electrophoresis, the gel was fixed (10% acetic acid/10% methanol for 30 min), dried onto filter paper, and subjected to autoradiography.

UV cross-linking

The binding reactions using oligonucleotides 1a/b and 2a/b as probes were scaled up five-fold and performed in duplicate. After incubation on ice, samples were subjected to UV irradiation by a pulsed Nd:YAG laser (5 nsec pulse at 266 nm, 60 mJ/pulse; Quanta-Ray, GCR14). One set of each duplicate reaction was subjected to electrophoresis through a standard band-shift gel. After electrophoresis, the gel was subjected to autoradiography and gel slices corresponding to the shifted bands on the autoradiograph were removed. The gel slices were placed in the sample wells of a discontinuous 10% SDS-polyacrylamide gel (Laemmli, 1970) along with the remaining set of UV treated binding reactions and a pre-stained protein size marker (BRL). After electrophoresis, the gel was fixed, dried and subjected to autoradiography as described above.

Part 1: Isolation and characterization of a NADP-dependent

glutamate dehydrogenase (GDH) gene

Results

Cloning of the GDH gene

A partial cDNA clone encoding a polypeptide homologous to eukaryotic glutamate dehydrogenases was fortuitously isolated during an attempt to clone the *Giardia lamblia* gene encoding the protein equivalent to the *E. coli* L12 ribosomal protein. Most characterized eukaryotic L12 equivalent genes contain the sequence MGFGLFD at their carboxy termini (Newton *et al.*, 1990). A 21-mer oligonucleotide complementary to mRNA encoding this heptapeptide (oLW72) was used along with an oligonucleotide *Xba*I-d(C)$_{14}$ as primers in a PCR reaction using poly(G) tailed cDNA as template (see Materials and Methods). One amplification product, cloned into the *Hinc*II site of pUC13, was 385 nucleotides in length. This clone, designated pJYPCR10.5, was found to encode a polypeptide that exhibits 77% identity over a region of 69 amino acid residues to the NADP-specific glutamate dehydrogenase of *Neurospora crassa*. I chose to characterize this GDH-like gene more fully because very few *Giardia* genes have been studied in detail.

The cDNA clone pJYPCR10.5 was used to screen a λEMBL3 *G. lamblia* genomic library. Eight positive clones were identified; all were found to contain a 2.4 kb *Pvu*II fragment that hybridized to the pJYPCR10.5 cDNA fragment. This 2.4 kb fragment was subcloned in both orientations into the *Sma*I site of pGEM3Zf(-) to give pJY1 and pJY11. The sequence of a 1691 nucleotide long portion of the 2.4 kb fragment, from an internal *Hind*III site to the downstream *Pvu*II site, was determined (Figure 5). The sequenced region was found to contain an open reading frame specifying a 449 amino acid long

```
HindIII         20              40              60              80              100             120
AAGCTTCACGTCGTCGTTCTCGTCTCCGCTGTCTGCATCCTTTACTTATCGCTTATAGAATGTCCACAAACTAAAAGTATACCGGTCAAACTTATTCTGAGGGTTACGGAACAAGTGTGC

                140             160             180             200             ••••            240
GCGTAGAGAGCCGGGGGCTCCCGAGGACCTTCGGGGGAGATGGCTGTTCTTGAAAAGCTGACCACAAATAACGCCTTTAATTACAGGCGCCCCAGATTTTAAAATGCCTGCCCAGACGAT
                                                                                                M  P  A  Q  T  I

        SacI    260             280             300   oJY 1           320             340             360
CGAGGAGCTCATCGCGGTGATCAAGCAGAGGGATGGCCACATGACGGAGTTCCGCCAGGCCGTCGAGGAGGTCGTGGACTCCCTCAAGGTTATCTTCGAGCGCGAGCCGAAGTATATCCC
E  E  L  I  A  U  I  K  Q  R  D  G  H  M  T  E  F  R  Q  A  U  E  E  U  U  D  S  L  K  U  I  F  E  R  E  P  K  Y  I  P

                380             400             420             Hincll          460             480
AATCTTCGAGAGGATGCTTGAGCCGGAGCGCGTCATCATCTTCCGCGTGCCCTGGATGGATGACGCTGGACGCATCAACGTCAACCGCGGCTTCCGTGTCCAGTACAACTCTGCTCTCGG
I  F  E  R  M  L  E  P  E  R  U  I  I  F  R  U  P  W  M  D  D  A  G  R  I  N  U  N  R  G  F  R  U  Q  Y  N  S  A  L  G

                500             520             540             560             580    oLW 72    600
CCCCTACAAGGGTGGCCTCCGCTTCCACCCCTCTGTCAATCTTTCGATTCTCAAGTTCCTCGGTTTCGAGCAGATCCTGAAGAACTCCCTCACCACGCTCCCGATGGGCGGCGGCAAGGG
P  Y  K  G  G  L  R  F  H  P  S  U  N  L  S  I  L  K  F  L  G  F  E  Q  I  L  K  N  S  L  T  T  L  P  M  G  G  G  K  G

__ _            620             640             660             680             700             720
CGGCTCCGACTTTGACCCAAAGGGCAAGTCCGACAACGAGGTCATGCGCTTCTGCCAGTCCTTCATGACCGAGCTCCAGAGGCACGTCGGCGCCGACACTGACGTTCCTGCCGGCGACAT
G  S  D  F  D  P  K  G  K  S  D  N  E  U  M  R  F  C  Q  S  F  M  T  E  L  Q  R  H  U  G  A  D  T  D  U  P  A  G  D  I

                740             760             780             800             820             840
CGGCGTCGGCGCCCGCGAGATCGGGTACCTGTACGGACAGTACAAGCGCCTGAGGAACGAGTTCACAGGCGTCCTCACAGGCAAGAACGTCAAGTGGGGCGGGTCTTTCATCAGGCCGGA
G  U  G  A  R  E  I  G  Y  L  Y  G  Q  Y  K  R  L  R  N  E  F  T  G  U  L  T  G  K  N  U  K  W  G  G  S  F  I  R  P  E

                860             880             900             920             940             960
GGCCACGGGCTATGGCGCTGTCTACTTCCTGGAGGAGATGTGCAAGGACAACAACACTGTGATCAGGGGTAAGAACGTCCTTCTTTCTGGCTCCGGCAACGTTGCCCAGTTTGCTTGCGA
A  T  G  Y  G  A  U  Y  F  L  E  E  M  C  K  D  N  N  T  U  I  R  G  K  N  U  L  L  S  G  S  G  N  U  A  Q  F  A  C  E

                980             1000            1020            1040            1060            1080
GAAGCTCATTCAGCTCGGCGCAAAGGTCCTCACCTTCTCAGACTCCAACGGGACCATTGTCGACAAGGACGGGTTCAACGAGGAGAAGCTGGCCCACCTCATGTACCTCAAGAACGAGAA
K  L  I  Q  L  G  A  K  U  L  T  F  S  D  S  N  G  T  I  U  D  K  D  G  F  N  E  E  K  L  A  H  L  M  Y  L  K  N  E  K

                1100            1120            1140            1160            1180            1200
GCGTGGGCGTGTTTCCGAGTTCAAGGACAAGTATCCCAGCGTCGCGTACTACGAGGGCAAGAAGCCTTGGGAGTGCTTCGAGGGCCAGATGGATTGCATCATGCCTTGCGCCACTCAGAA
R  G  R  U  S  E  F  K  D  K  Y  P  S  U  A  Y  Y  E  G  K  K  P  W  E  C  F  E  G  Q  M  D  C  I  M  P  C  A  T  Q  N

                1220            1240            1260            1280            1300            1320
CGAGGTTTCCGGGGACGATGCGACGCGCCTGGTCGGCCTCGGCCTCAAGTTCGTGGCCGAGGGTGCGAACATGCCCTCCACGGCAGAGGCTGTTCACGTCTATCACGCCAAGGGCGTGAT
E  U  S  G  D  D  A  T  R  L  U  G  L  G  L  K  F  U  A  E  G  A  N  M  P  S  T  A  E  A  U  H  U  Y  H  A  K  G  U  M

                1340            1360            1380   EcoRI    1400            1420            1440
GTACGGACCCGCCAAGGCCAGCAACGCGGGCGGTGTCTCTGTCTCCGGCCTCGAGATGTCCCAGAATTCCGTGAGGCTCCAGTGGACGGCTGAGGAGGTCGACCAGAAGCTCCGCGGCAT
Y  G  P  A  K  A  S  N  A  G  G  U  S  U  S  G  L  E  M  S  Q  N  S  U  R  L  Q  W  T  A  E  E  U  D  Q  K  L  R  G  I

                1460            1480            1500            1520            1540            1560
CATGAGGGGCATCTTCGTCGCCTGCCGCGACACTGCCAAGAAGTATGGGCACCCCAAGAACTACCAGATGGGCGCGAACATCGCCGGGTTCCTGAAGGTCGCAGACTCTATGATCGAGCA
M  R  G  I  F  U  A  C  R  D  T  A  K  K  Y  G  H  P  K  N  Y  Q  M  G  A  N  I  A  G  F  L  K  U  A  D  S  M  I  E  Q

                1580            oooo            1620            1640            1660            NarI
GGGCTGCGTGTGAGGGCTGAAGTGAATATTTACCTTTTCCGGGTTCGACGGGCGTGCTTGACGATGCGGAAGTCGCTGTCCGTTCCAACGGCTATCTGCCCCAGGACCCCCATCCCGGCG
G  C  U
```

        PvuII
CCGATCAGCTG

Figure 5: Nucleotide sequence of the *Giardia* GDH gene

The nucleotide sequence of the 1691 nucleotide long *Hind*III-*Pvu*II fragment is illustrated. The location of the indicated restriction enzyme sites are: *Hind*III (1), *Sac*I (245), *Hinc*II (441), *Eco*RI (1384), *Nar*I (1677) and *Pvu*II(1686). These are the sites used to generate fragments for the S1 nuclease protection experiments and other experiments discussed in the main text. The sequences complementary to oligonucleotides oWL72 and oJY1 are indicated by overlines; the interruptions in the oWL72 overline represent nucleotides that are not complementary to the GDH sequence. The position of 5'(•) and 3'(o) transcript end sites are indicated. A putative polyadenylation signal is underlined at positions 1581-1586.

GDH-like protein. The protein-coding portion of this gene has a G+C content of 61%. The oligonucleotide oLW72, used in the initial PCR amplification, matches the GDH sequence at 14 out of 21 nucleotide positions located between 584-604. The other end of the pJYPCR10.5 cDNA clone corresponds to genomic nucleotide position 220 and the cDNA sequence agrees precisely with the genomic sequence between positions 220 and 583.

The insert from the clone pJYPCR10.5 was also used to probe a *G. lamblia* cDNA library in λgt10. The insert of one of the hybridizing cDNA clones was subcloned into the *Eco*R1 site of pGEM3Zf(-), designated pJYc4-7, for sequencing. The cDNA sequence extended from nucleotide position 234 to the *Eco*RI site at 1389 and was colinear and identical in sequence to the genomic clone, implying that there are no introns within this region of the *Giardia* GDH-like gene.

Characterization of the GDH gene transcripts

The 5' ends of the transcripts derived from this open reading frame were characterized by primer extension and S1 nuclease protection. A 441 bp *Hind*III-*Hinc*II fragment overlapping the 5' portion of the open reading frame was 5' end-labeled in the (-) DNA strand at nucleotide position 442. Following its hybridization to total *Giardia* RNA and S1 nuclease digestion of the heteroduplex, a 222 nucleotide long doublet, representing the protected product, was observed (Fig. 6B). This places the 5' mRNA end in the vicinity of nucleotide position 221, three nucleotides in front of the ATG translation initiation codon.

This result was confirmed using primer extension analysis using a 20-mer oligonucleotide (oJY1), complementary to the (+) DNA strand of the genomic sequence between nucleotide position 299-319, as presented in figure 5. Using total *Giardia* RNA as template, a major extension product 98 nucleotides in length was observed; minor products one to three nucleotides longer were also apparent (Fig. 6A). These 5' end sites
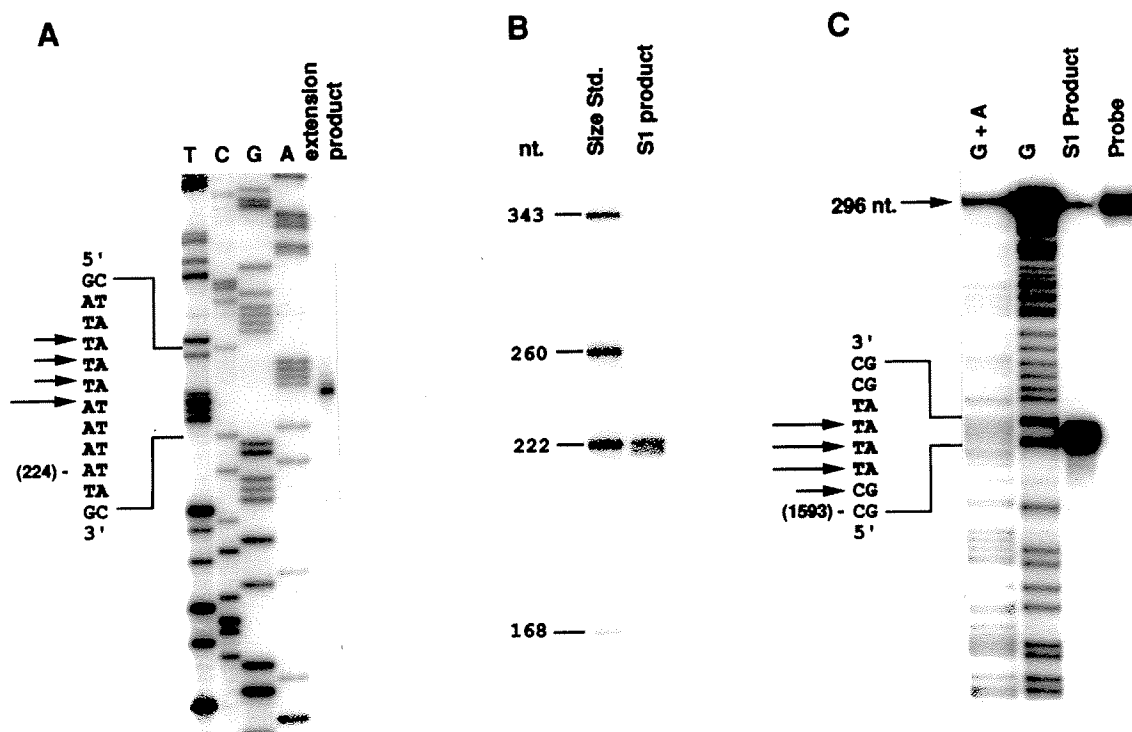
Figure 6: Mapping of the 5' and 3' end sites of the GDH mRNA transcript

Primer extension or S1 nuclease protection assays were used to locate the 5' and 3' ends of the GDH mRNA. Radioactive products of these assays were separated on denaturing polyacrylamide gels and subjected to autoradiography. (A) Primer extension analysis using 5' end-labeled oJY1 oligonucleotide and total *Giardia* RNA was carried out as described in the "Materials and Methods". The length of the product was estimated using a DNA sequence ladder generated using the same oJY1 oligonucleotide as a sequencing primer. The major end site (large arrow) occurs three nucleotides in front of the ATG translation initiation codon. Minor end sites (small arrows) occur four to six nucleotides in front of the ATG codon. The A of the ATG codon, at position 224, is indicated within the sequence shown. (B) Nuclease S1 protection by mRNA of an end-labeled DNA probe. A 441 nucleotide long *Hind*III-*Hinc*II fragment was 5' end- labeled in the (-) DNA strand at the *Hinc*II site at position 443. Following hybridization to total *Giardia* RNA and S1 digestion, a protected product of about 222 nucleotides in length was observed. The molecular lengths of size standards in nucleotides are indicated on the left. (C) Nuclease S1 protection by mRNA of an end-labeled DNA probe. A 296 nucleotide long *Eco*RI-*Nar*I fragment was labeled at the 3' end of the (-) DNA strand at the *Eco*RI site at positions 1385-1388. The molecular length standard was generated by employing the G and G+A Maxam-Gilbert sequencing reaction on the same end-labeled fragment. The protected product 207-210 nucleotides in length positions the 3' transcript end sites near nucleotide positions 1594-1597. Nucleotide position 1593 is indicated within the sequence shown.

occur at nucleotide positions 218-221 in the genomic DNA sequence, and correspond to the 5' end sites determined by S1 nuclease protection. Furthermore, they also correspond to the 5' end of the original pJYPCR10.5 cDNA insert at nucleotide position 221.

The 3' transcript end sites were also located by S1 nuclease protection (Fig. 6C). The 296 bp EcoRI-NarI fragment overlapping the 3' end of the open reading frame was 3' end-labeled in the (-) DNA strand at the EcoRI site (nucleotide position 1285-8). Following hybridization to total Giardia RNA and digestion with S1 nuclease, protected fragments of 207-210 nucleotides in length were observed. This places the 3' transcript end site at nucleotide positions 1594-1597, preceded by a putative Giardia polyadenylation signal sequence AGTGAA at nucleotide position 1478-1483.

These results demonstrate that the major transcript of the GDH gene excluding the poly(A) tail is about 1380 nucleotides in length. The major 5' end site is located only three nucleotides in front of the putative ATG translation initiation codon and the transcript is probably polyadenylated at a site that is 21-24 nucleotides beyond the translation termination site. The 5' untranslated leader is not extended by transsplicing as indicated by primer extension analysis; presumably, the observed 5' end site represents the position of transcription initiation.

Genomic Southern hybridization

To determine if the Giardia haploid genome contains other copies of the cloned GDH gene or related genes, restriction enzyme digests of genomic DNA were probed with the 2.4 kb PvuII fragment under high stringency conditions in Southern hybridizations (Fig. 7). Only single hybridizing bands, all greater than 6 kb in length, were evident in BamHI, BglII, EcoRI, and PstI digests. These fragments span a region greater than 9 kb in length and encompass the entire GDH gene, which implies that the haploid genome probably contains a single GDH-like sequence. The partial restriction enzyme map of
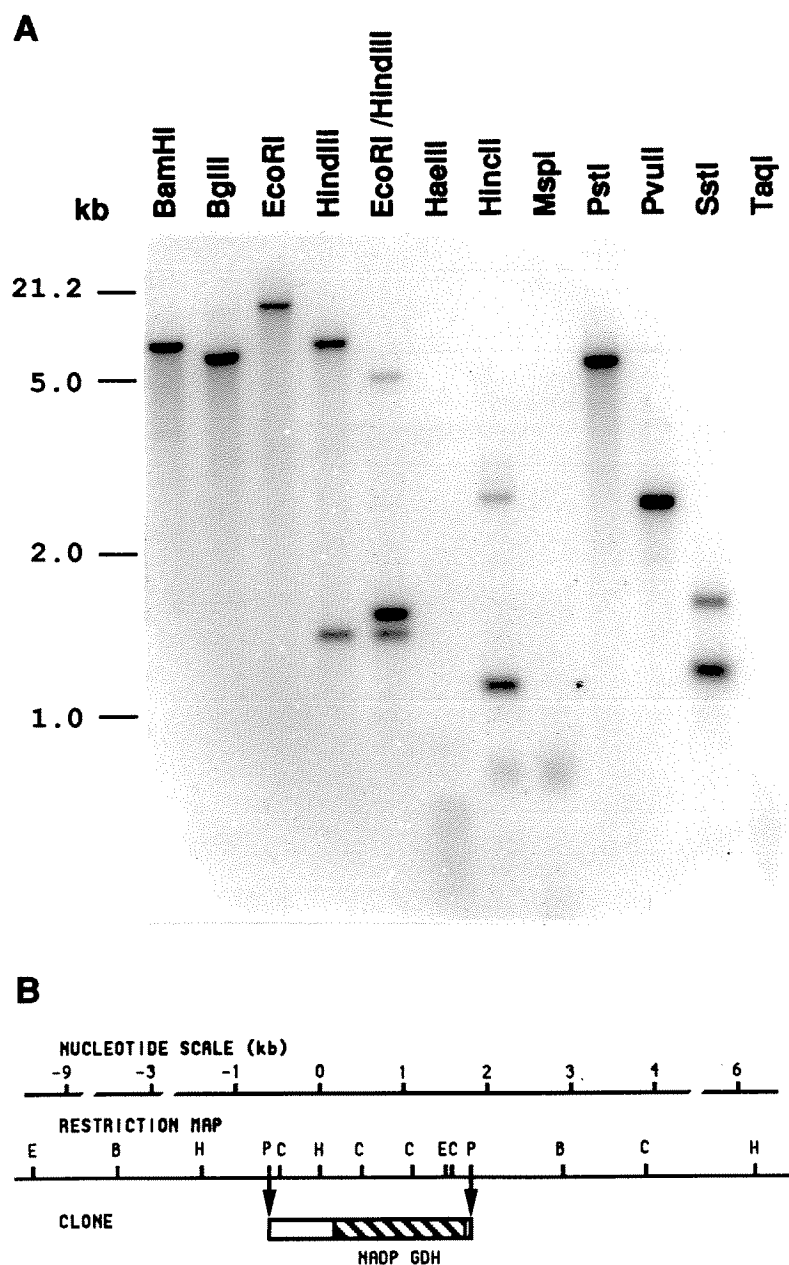
Figure 7: **Genomic Southern hybridization to identify fragments encoding NADP-GDH or related proteins**

(A) Genomic *G. lamblia* DNA was digested with different restriction enzymes and separated on a 0.8% agarose gel. Following transfer to nylon membrane, the DNA was probed with the 2.4 kb *Pvu*II fragment from the pJY11 clone. (B) A partial restriction map of the genomic DNA containing the GDH gene was constructed from the Southern hybridization results as illustrated and from other hybridization results using the insert from pJYPCR10.5 as the probe (data not shown). The nucleotide scale is presented above, the restriction map in the middle, and the position and polarity of the GDH gene on the 2.4 kb *Pvu*II fragment on the bottom. Restrictions sites are as follows: E, *Eco*RI; B, *Bam*HI; H, *Hind*III; P, *Pvu*II; C, *Hinc*II.

the region, along with the position and orientation of the GDH gene is included in figure 7. It is interesting to note that only a single EcoRI fragment was detected in these experiments in spite of the fact that this enzyme cuts once asymmetrically within the 2.4 kb PvuII probe. I suspect that the second fragment is very large because, by empirical observations, Giardia DNA appears to be extremely deficient in EcoRI recognition sites. This large fragment may transfer inefficiently during the blotting process or may hybridize poorly because it overlaps the probe by only 300 nucleotides, and therefore has escaped detection. The other end of the detectable EcoRI fragment has been mapped to a position about 9 kb upstream of the GDH coding region.

Enzyme activity

The presence of mRNA transcripts from the putative NADP-dependent GDH gene predicts that Giardia cell extracts should exhibit this enzymatic activity. In the presence of the substrates α-ketoglutarate and NH_4Cl, a crude cell lysate exhibited a specific activity of 210 ± 20 mU per mg of protein. No significant NAD-dependent GDH activity was detected in either the direction of reductive amination or oxidative deamination.

The fungi Neurospora crassa and Saccharomyces cerevisiae and the protozoan Trypanosoma cruzi have, in addition to a class I NADP-dependent GDH, a class II NAD-dependent GDH (Kinnaird and Fincham, 1983; Cazzulo, 1984; Moye et al., 1985; Nagasu and Hall, 1985; Cazzulo et al., 1988; Vierula and Kapoor, 1989; Miller and Magasanik, 1990). A clone of the gene encoding the S. cerevisiae NAD-GDH was used to probe the Giardia genome by Southern blot analysis to identify the homologous gene. No specific hybridization was detected. Together with the enzyme assays, these results indicate that Giardia has the single NADP-dependent GDH.

Sequence alignments

The amino acid sequence specified by the *Giardia* GDH gene was aligned over its entire length to the amino acid sequence of the NADP-dependent GDHs from *Neurospora crassa* (a fungus) and *Escherichia coli* (a eubacterium) (Figure 8). The alignments are disrupted by 11 separate gaps presumed to reflect deletion or insertion events during the course of evolution. Of the 11 gap positions, seven are shared jointly between *E. coli* and *Giardia* and only three are shared by *Neurospora* and *Giardia* (Table I). *Neurospora* and *E. coli* have no gaps in common. The remaining gap at 310-312, because it is different in each sequence, is phylogenetically uninformative.

It is also surprising to see that in pairwise comparisons at common positions, the *Giardia* sequence is 55% and 58% identical to the *N. crassa* and *E. coli* sequence, respectively (Table I). The *E. coli* sequence exhibits 56% identity with the *Neurospora* sequence, and nearly 47% of the commonly shared amino acid positions are identical in all three sequences. These amino acid identity data argue that the *Giardia* GDH protein sequence is at least as closely related to the eubacterial sequence as it is to the fungal sequence. This conclusion is further supported by the distribution of the gap positions in the alignment of the proteins.

```
              10          20          30
Gla    M P A Q T I E E L I A V I K Q R D G H M T E F R Q A V E E V V D S L • • K V I
Ncr                          S N L P S E P - - E - - Y K[]L A Y T - E N S S L
Eco    M D Q T Y S L - S F L N H V Q K - - P N Q - - - A - - - R - - M T T - • • W P F

              40          50          60          70
Gla    F E R E P K Y • • I P I F E R M L E P E R V I I F R V P W M D D A G R I N V N R
Ncr    - Q K H - E - • • R T A L T V A S I - - - - - Q - - - V - E - - N - N V Q - - -
Eco    L - Q N - - - R Q M S L L - - L V - - - - - - Q - - - V - V - - R N Q - Q - - -

              80          90          100         110
Gla    G F R V Q Y N S A L G P Y K G G L R F H P S V N L S I L K F L G F E Q I L K N S
Ncr    - Y - - - F - - - - - - - - - - - L - - - - - - - - - - - - - - - - - F - - A
Eco    A W - - - F S - - I - - - - - - M - - - - - - - - - - - - - - - - - T F - - A

              120         130         140         150
Gla    L T T L P M G G G K G G S D F D P K G K S D N E V M R F C Q S F M T E L Q R H V
Ncr    - - G[]S - - - - - - - A - - - - - - - - - A - I R - - - C A F - A - - H K - I
Eco    - - - - - - - - - - - - - - - - - - - - - E G - - - - - - - - L - - - - Y - - L

              160         170         180         190
Gla    G A D T D V P A G D I G V G A R E I G Y L Y G Q Y K R L R N E F T G V L T G K N
Ncr    - - - - - - - - - - - - - - G - - - - - M F - A - R K A A - R - E - - - - - - G
Eco    - - - - - - - - - - - - - G - - V - F M A - M M - K - S - N T A C - F - - - G

              200         210         220         230
Gla    V K W G G S F I R P E A T G Y G A V Y F L E E M C K • D N N T V I R G K N V L L
Ncr    L S - - - - L - - - - - - - - - L - - V V G H - L E Y S G A G S Y A - - R - A -
Eco    L S F - - - L - - - - - - - - - L - - - T - A - L - • R H G M G F E - M R - S V

              240         250         260         270
Gla    S G S G N V A Q F A C E K L I Q L G A K V L T F S D S N G T I V • • D K D G F N
Ncr    - - - - - - - - Y - A L - - I E - - - T - V S L - - - K - A L - A T G E S - I T
Eco    - - - - - - - - Y - I - - A M E F - - R - I - A - - - S - - V - • • - E S - - T

              280         290         300         310
Gla    E E K L A H L M Y L K N E K R G R V S E F K D K Y P S V A Y Y E G K K P W E C F
Ncr    V - D I N A V - A I - • - A - Q S L T S - Q H A • G H L K W I - - A R - - L • H
Eco    K - - - - R - I E I - A S R D - - - A D Y A K E F G • L V - L - - Q Q - - • • •

              320         330         340         350
Gla    E G Q M D C I M P C A T Q N E V S G D D A T R L V G L G L K F V A E G A N M P S
Ncr    V - K V - I A L - - - - - - - - - K E E - E G - L A A - C - - - - - - S - - G C
Eco    S L P V - I A L - - - - - - - - - V - A - H Q - I A N - V - A - - - - - - - - T

              360         370         380
Gla    T A E A V H V Y H • • • • • • • A K G V M Y G P A K A S N A G G V S V S G L E M
Ncr    - L - - I E - F E N N R K E K K G E A - - - A - G - - A - C - - - A - - - - - -
Eco    - I - - T E L F Q • • • • • • • Q A G - L F A - G - - A - - - - - A T - - - - -

              390         400         410         420
Gla    S Q N S V R L Q W T A E E V D Q K L R G I M R G I F V A C R D T A K K Y G H P K
Ncr    A - - - Q - - N - - Q A - - - E - - K D - - K N A - F N G L N - - - T - V E A A
Eco    A - - A A - - G - K - - K - - A R - H H - - L D - H H - - V E H G G E • - E Q T

              430         440
Gla    N Y Q M • • • • • • G A N I A G F L K V A D S M I E Q G C V
Ncr    E G E L P S L V A - S - - - - - V - - - Q A - H D - - D W W S K N
Eco    - - V Q • • • • • • - - - - - - V - - - - A - L A - - V I
```

Figure 8: **Glutamate dehydrogenase protein sequence alignments**
The NADP-glutamate dehydrogenase amino acid sequences from *Neurospora crassa* (Ncr) and *Escherichia coli* (Eco) are aligned to the predicted *Giardia lamblia* (Gla) protein sequence. Positions in *N. crassa* and *E. coli* sequences that are identical to the *G. lamblia* amino acid are indicated by dashes (-); positions of deletions or insertions necessary to maintain optimal alignment are indicated by filled circles (•). The amino acid numbering is based on the *G. lamblia* protein sequence. The amino acids that are specified by codons which are interrupted by introns in the *N. crassa* gene are boxed.

Table I: Comparison of homologue GDH proteins

Comparisons[1]	Common Amino Acid Positions[2]	Amino Acid Identities[3]	Gaps in Common[4]
Gla Ncr	432	239 (55.3 %)	3
Gla Eco	442	255 (57.7 %)	7
Ncr Eco	428	240 (56.1 %)	0
Gla Ncr Eco[5]	428	200 (46.7 %)	--

[1]Abbreviations are: Gla, *Giardia lamblia*; Eco, *Escherichia coli*; Ncr, *Neurospora crassa*.

[2]Common amino acid positions represent the number of shared amino acid positions in a pairwise alignment.

[3]Amino acid identities are the number of positions where identical amino acids occur in the pairwise alignments. The percentages of amino acid identities are in brackets.
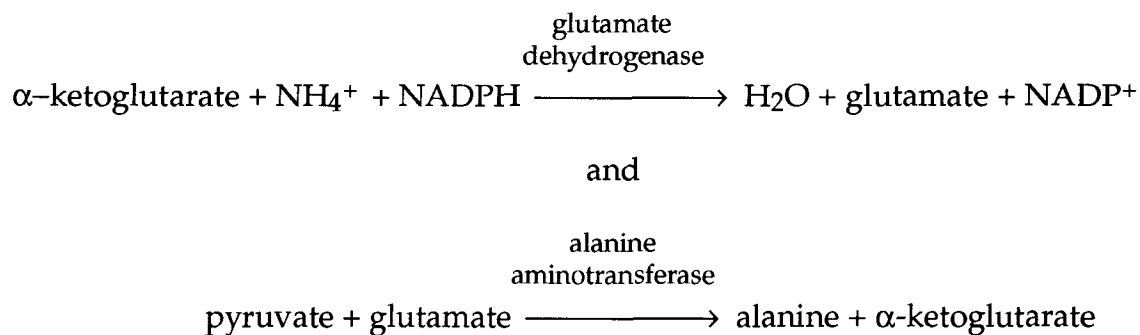
[4]Gaps in common are those that occur at the same position in two of the sequences that are required to maintain overall alignment with the third sequence.

[5]The bottom row illustrates the common amino acid positions and the amino acid identities shared between the three sequences.

Discussion

<u>Function of the GDH protein</u>

In both eukaryotic and prokaryotic organisms, the primary role of NADP-dependent glutamate dehydrogenase is normally to produce glutamate for the support of protein synthesis. In contrast, anaerobic *Giardia* utilize this enzyme almost exclusively to maintain a redox balance within the intracellular NAD(P) pool (*Paget et al., 1993*). In the anaerobic metabolism of *Giardia*, glucose or other carbohydrates are converted to pyruvate via the Embden-Meyerhoff pathway with the concomitant reduction of NAD(P) (Lindmark, 1980; Jarroll *et al.*, 1989). Pyruvate is then converted to a variety of end products including acetate, ethanol, CO_2, and alanine (Fig. 9) (*Paget et al., 1993*). The conversion to alanine utilizes two coupled reactions:

$$\alpha\text{-ketoglutarate} + NH_4^+ + NADPH \xrightarrow{\text{glutamate dehydrogenase}} H_2O + glutamate + NADP^+$$

and

$$pyruvate + glutamate \xrightarrow{\text{alanine aminotransferase}} alanine + \alpha\text{-ketoglutarate}$$

A number of unicellular eukaryotes (fungi and protozoa) possess a second GDH that utilizes NAD rather than NADP as coenzyme (Cazzulo *et al.*, 1988; Stevens *et al.*, 1989; Vierula and Kapoor, 1989; Miller and Magasanik, 1990). This activity is associated with the conversion of glutamate to α-ketoglutarate for catabolic purposes. Southern blot analysis of *Giardia* genomic DNA using the *S. cerevisiae* NAD-GDH as a probe did not detect the presence of a related sequence. Furthermore, no NAD-dependent interconversion between α-ketoglutarate and glutamate was evident in *Giardia* cell extracts; only the NADP-dependent activity was detectable.
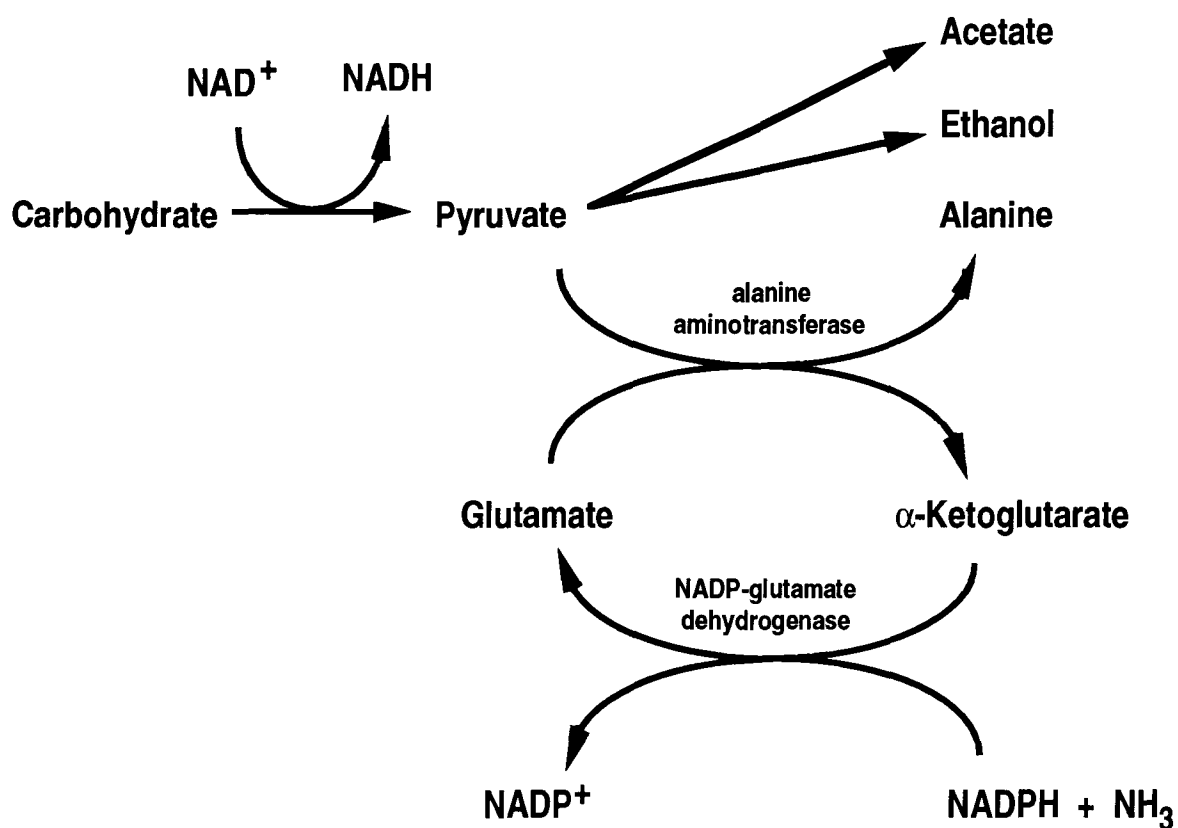
Figure 9: **The role of NADP-GDH in** *Giardia* **metabolism.**
Giardia uses anaerobic glycolysis to generate pyruvate which is then converted to three separate end products, acetate, ethanol, and alanine. The role of NADP-glutamate dehydrogenase and alanine aminotransferase in the conversion of pyruvate to alanine is illustrated.

Evolutionary relationships

The deduced amino acid sequence of the *Giardia* GDH protein exhibits substantial sequence identity to numerous fungal and eubacterial NADP-GDH sequences. Figure 8 illustrates the amino acid sequence alignment of the *Giardia* protein to the *N. crassa* and the *E. coli* proteins. Comparison of amino acid gap positions and amino acid identities indicate that the *Giardia* GDH is no more similar to the *Neurospora* enzyme than it is to the *E. coli* enzyme. The same conclusion is obtained when the *Giardia* sequence is compared to other fungal or other eubacterial NADP-GDH sequences (data not shown). If inheritance of the NADP-GDH gene has been entirely vertical from a common ancestor, these protein comparisons provide strong support for the hypothesis that *Giardia* diverged very early from the main eukaryotic lineage. However, these sequence comparisons were performed on GDHs from just three organisms by rudimentary analysis. Later studies by Benachenhou-Lahfa *et al.*, (1993) using the *Giardia* GDH along with class I GDHs from 20 other organisms in a more extensive analysis, suggested that the class I GDHs can be further sub-divided into two families with the GDHs from *Giardia, Neurospora,* and *E. coli.* belonging to the same family. Depending on the portion of the GDH sequence and the method of tree reconstruction used, the *Giardia* sequence either groups with the eubacteria or the fungi sequences. This and other discrepancies in the trees derived from the GDH sequence comparisons by Benachenhou-Lahfa *et al.*, suggest that the evolution of GDHs may be unusual and more complex than expected.

Structure and expression of the GDH gene

The *N. crassa* NADP-GDH gene contains two spliceosomal introns that disrupt the gene within the codons at alignment positions 29 and 119 (Figure 8). In contrast, the *Giardia* genomic and cDNA clones (covering 87% of the coding region) are colinear and identical. Together with S1 nuclease protection experiments, these results demonstrate

that there are no introns in the *Giardia* GDH gene and that at least 87%, and probably the entire RNA transcript, is unaffected by post-transcriptional RNA editing. Features such as introns, RNA editing and transsplicing have yet to be detected in *Giardia*. However, since the number of fully characterized genes and transcripts is small, it is still not possible to make general statements regarding the evolutionary origins of these features and their frequency and distribution within *Giardia*.

There are only three to six nucleotides in front of the AUG translation initiation codon on the *Giardia* GDH mRNA. Primer extension, S1 nuclease protection and RNA-PCR experiments all indicate that this short 5' transcript leader is not extended by a transsplicing event. The DNA sequence overlapping the 5' transcript end site and the translation initiation codon is AT-rich (5'...AGATTTTAAA<u>ATG</u>...'3) and provides a striking contrast to the generally GC-rich *Giardia* coding sequences. Other *Giardia* mRNA transcripts have also been shown to have 5' leaders of one to six nucleotides in length (reviewed by Adam, 1991). For a number of these, the 5' transcript end sites fall within AT-rich sequences which exhibit a high degree of sequence similarity to this GDH AT-rich sequence. In part 2 of this report, I present evidence that this motif is common to other *Giardia* genes and contains a specific binding site for a *Giardia* nuclear protein.

It seems likely that *Giardia* uses a unique mechanism for translation initiation. The ribosome scanning model for translation initiation in other eukaryotes requires the small ribosomal subunit to scan along the mRNA and identify the AUG initiation codon within an appropriate context sequence (Kozak, 1989). Because the AUG initiation codon is so close to the 5' end of the mRNA, ribosome scanning may not be required for translation initiation of *Giardia* mRNAs.

The sequence AGTPuAA has been found in the 3' untranslated region of all *Giardia* genes that have been examined. This sequence has been proposed to function as a polyadenylation signal because of (i) its location relative to the translation termination

codons and the site of polyadenylation, and (ii) its similarity to the eukaryotic consensus polyadenylation signal (5'...AATAAA...3') (Holberton *et al.*, 1988; Peattie *et al.*, 1989). A sequence (5'...AGTGAA...3') matching the putative *Giardia* polyadenylation signal occurs in the GDH transcript eight nucleotides downstream from the translational termination codon. Nuclease S1 protection experiments suggest that the polyadenylation site is located eight to 11 nucleotides beyond this putative polyadenylation signal.

The presence of a poly(A) tail on the GDH mRNA has not been verified directly in this study. However, the λgt10 library from which the GDH cDNA clone was isolated was constructed using oligomer $d(T)_n$ as first strand primer. Therefore, the ability to isolate a GDH clone from this library implies that the GDH mRNA is polyadenylated.

During the construction of the cDNA library, the *Eco*RI sites within the cDNA products were protected by *Eco*RI methylase before *Eco*RI linkers were added to the ends. The ligation products were then cut with *Eco*RI and ligated into the λ vector. The finding that one end of the GDH cDNA clone from this library contains a naturally occurring *Eco*RI site within the GDH gene suggests that this site was not protected by the methylase during the construction of the library.

In summary, the results indicate that *Giardia* contains a single class I GDH gene. Similar to *Giardia* genes characterized previously, the GDH gene is GC-rich (61%), and its transcripts have short (three to six nucleotides) untranslated 5' sequences. A sequence matching a putative *Giardia*-specific poly(A) signal is found at in the 3' flanking region of the gene. The sites of poly(A) addition on the GDH mRNA, located by S1 nuclease protection, is consistent with the usage of the putative poly(A) signal. Comparisons of amino acid identities suggests that the *Giardia* GDH is no more closely related to homologous fungal enzymes than it is to eubacterial enzymes. The failure to identify introns, transsplicing or RNA editing in the transcript of the GDH gene may indicate that these molecular features are rare or absent in the *Giardia* lineage.

Part 2: Identification and analysis of conserved sequence

elements in the GDH gene promoter

Results

Identification of conserved sequence motifs

The upstream sequences from eleven *Giardia* genes, including the NADP-dependent glutamate dehydrogenase gene (GDH), are aligned in Fig. 10. It is evident from the alignment that core promoter elements such as the TATA-box ($5'$TATA(A/T)AAT$^{3'}$) and the pyrimidine-rich initiator element $5'$(C/T)A(C/T)TC(C/T)(C/T)(C/T)$^{3'}$ that are normally associated with RNA polymerase II promoters of higher eukaryotes are missing from these regions. However, two other conserved sequence motifs are found (Fig. 10). One is an AT-rich sequence element adjacent to the ATG translation initiation codon and encompasses the transcription start sites. This element is present in ten of the eleven sequences examined. The AT-richness of this element resembles the TATA-box but its location resembles the initiator sequence of higher eukaryotes.

The second conserved element is a CAAAT sequence which is referred to as the g-CAB element for the *Giardia* CAAT-box like sequence. This g-CAB motif shares limited sequence identity with the CAAT-box found in mammalian promoters which have the consensus sequence, GG(C/T)CAATCT. This g-CAB element, in its complete form in either orientations, CAAAT or GTTTA, is found at least once between nucleotides -50 and -25 in the promoter region of seven *Giardia* genes. In its tetranucleotide abbreviated forms, CAAA, AAAT, GTTT and TTTA, the g-CAB element occurs at least three times in nine of the eleven promoter sequences including once or twice within most of the longer AT-rich motifs.
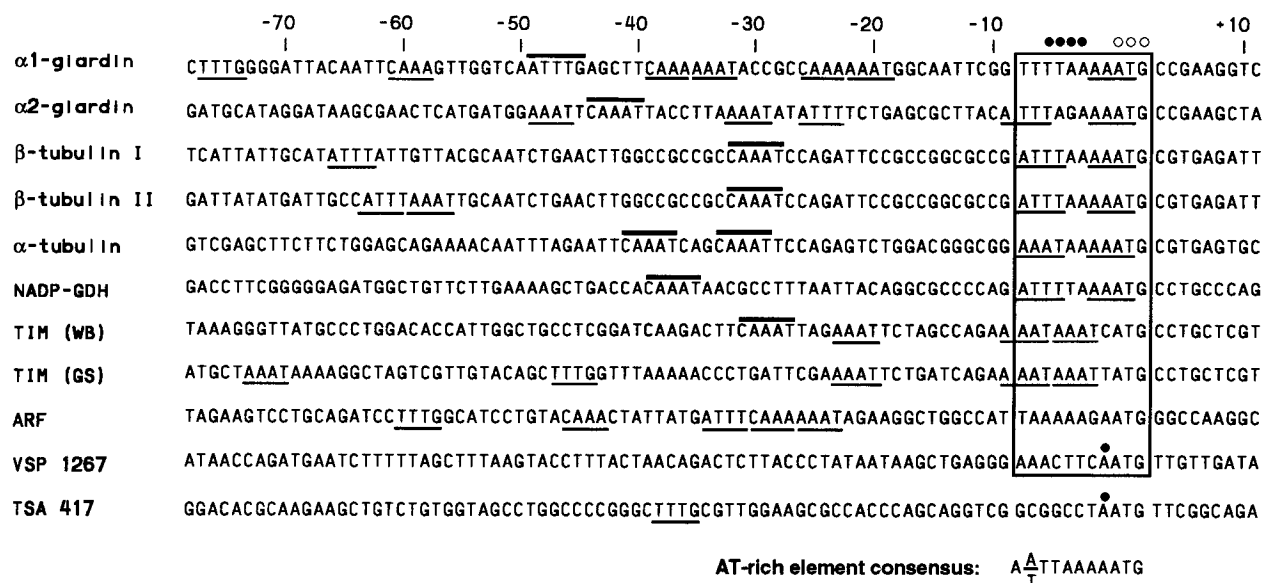
```
              -70       -60       -50       -40       -30       -20       -10              +10
               |         |         |         |         |         |         |     ••••  ooo    |
α1-giardin  CTTTGGGGATTACAATTCAAAGTTGGTCAATTTGAGCTTCAAAAAATACCGCCAAAAAATGGCAATTCGG|TTTTAAAAATG|CCGAAGGTC

α2-giardin  GATGCATAGGATAAGCGAACTCATGATGGAAATTCAAATTACCTTAAAATATATTTTCTGAGCGCTTACA|TTTAGAAAATG|CCGAAGCTA

β-tubulin I  TCATTATTGCATATTTATTGTTACGCAATCTGAACTTGGCCGCCGCCAAATCCAGATTCCGCCGGCGCCG|ATTTAAAAATG|CGTGAGATT

β-tubulin II GATTATATGATTGCCATTTAAATTGCAATCTGAACTTGGCCGCCGCCAAATCCAGATTCCGCCGGCGCCG|ATTTAAAAATG|CGTGAGATT

α-tubulin   GTCGAGCTTCTTCTGGAGCAGAAAACAATTTAGAATTCAAATCAGCAAATTCCAGAGTCTGGACGGGCGG|AAATAAAAATG|CGTGAGTGC

NADP-GDH    GACCTTCGGGGGAGATGGCTGTTCTTGAAAAGCTGACCACAAATAACGCCTTTAATTACAGGCGCCCCAG|ATTTTAAAATG|CCTGCCCAG

TIM (WB)    TAAAGGGTTATGCCCTGGACACCATTGGCTGCCTCGGATCAAGACTTCAAATTAGAAATTCTAGCCAGAA|AATAAATCATG|CCTGCTCGT

TIM (GS)    ATGCTAAATAAAAGGCTAGTCGTTGTACAGCTTTGGTTTAAAAACCCTGATTCGAAAATTCTGATCAGAA|AATAAATTATG|CCTGCTCGT

ARF         TAGAAGTCCTGCAGATCCTTTGGCATCCTGTACAAACTATTATGATTTCAAAAAATAGAAGGCTGGCCAT|TAAAAAGAATG|GGCCAAGGC

VSP 1267    ATAACCAGATGAATCTTTTTAGCTTTAAGTACCTTTACTAACAGACTCTTACCCTATAATAAGCTGAGGG|AAACTTCAATG|TTGTTGATA

TSA 417     GGACACGCAAGAAGCTGTCTGTGGTAGCCTGGCCCCGGGCTTTGCGTTGGAAGCGCCACCCAGCAGGTCG GCGGCCTAATG TTCGGCAGA
```

AT-rich element consensus: A$_T^A$TTAAAAATG

Figure 10: Identification of two conserved sequence elements in the upstream region of *Giardia* genes.

The 5' sequences of the *Giardia* genes presented in the alignment are as follows from top to bottom: α1- and α2 giardins (Peattie *et al.*, 1989; Alonso and Peattie, 1992), βI-, βII- and α-tubulins (Kirk-Mason *et al.*, 1989), NADP-dependent glutamate dehydrogenase (GDH) (Yee and Dennis, 1992), triosephosphate isomerase (TIM) from the WB and GS *Giardia* isolates (M. Mowatt, pers. comm.), ADP-ribosylation factor (ARF) (Murtagh *et al.*, 1992), cell surface antigens VSP 1267 (Mowatt *et al.*, 1992) and TSA 417 (Gillin *et al.*, 1990). The ATG translation initiation codon (o) and the sites of transcription initiation (•) are indicated. The transcripts of the first nine genes are initiated three to six nucleotides in front of the ATG codon and the transcripts from the last two genes initiate one nucleotide in front of the ATG codon. A conserved AT-rich DNA sequence element overlapping the 5' transcript end sites and the translation initiation codon is enclosed by the large box on the right. A second conserved sequence element g-CAB (with the sequence CAAAT or its inverted complement ATTTG) located between -25 and -50, is indicated by heavy overlines and tetranucleotide sequences related to g-CAB are underlined. The numbering of the nucleotides in the sequence is relative to the zero position assigned to the A of the ATG translation initiation codon.

The 5' sequences that appear most aberrant with respect to these motifs are from the variant-specific antigen genes TSA 417 and VSP 1267. The former completely lacks the AT-rich element and contains only a single copy of a tetranucleotide version of the g-CAB element whereas the latter contains an AT-rich element with limited identity to the consensus and no g-CAB-like elements. These observations suggest that the putative promoters of the variant-specific antigen genes differ from the putative promoters of genes encoding structural proteins and enzymes involved in central metabolism. In this study I have examined the role of the AT-rich and the g-CAB sequences in protein binding to the putative promoter region of the GDH gene.

Protein binding to double-stranded DNA

To investigate whether the AT-rich and the g-CAB motifs are recognition sites for protein binding, two probes were constructed to represent each of these motifs for use in band-shift assays (Fig. 11). The oligonucleotide duplex 1a/b represents the region of the GDH sequence between -13 and +9, and contains the AT-rich element. The oligonucleotide duplex 2a/b represents the region of the GDH sequence between -44 and -24, and contains the g-CAB element. Each of these oligonucleotide duplexes was radiolabeled and incubated with nuclear protein extracts prepared from *Giardia* (Fig. 12A, B). On a non-denaturing polyacrylamide gel, both probes in the presence of nuclear extract exhibited two bands with retarded mobilities relative to the mobility of the free probes. The two complexes, referred to as the upper and lower bands, represent specific binding of the nuclear proteins to the probes. Formation of these complexes is inhibited by the addition of an excess of unlabeled probe and not by the addition of the oligonucleotide duplex G4-a/b that contained the binding site for the yeast GAL4 protein (Johnston and Davis, 1984). (See Table II for oligonucleotide sequences). Surprisingly, oligonucleotides 1a/b and 2a/b, containing the AT-rich and the g-CAB motifs respectively, can also cross-compete with each other with
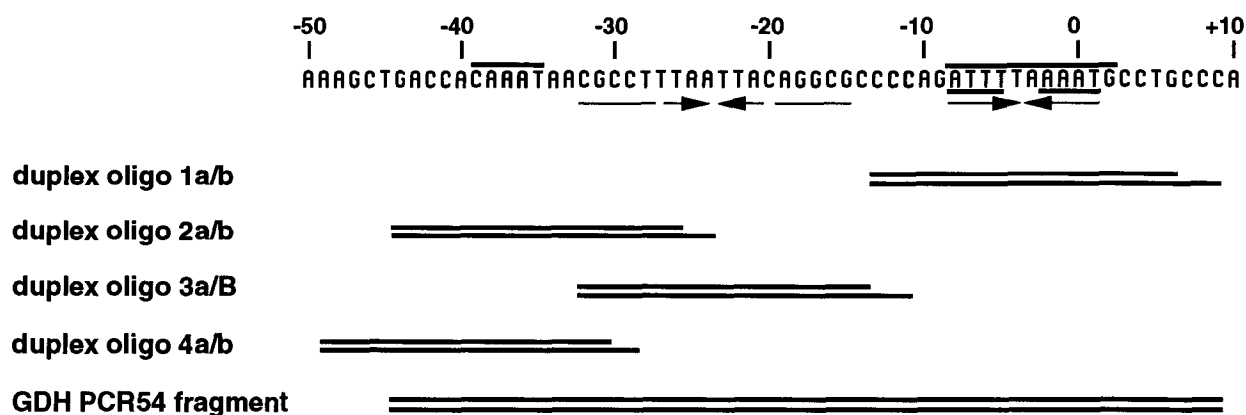
```
        -50         -40         -30         -20         -10          0          +10
         I           I           I           I           I           I            I
        AAAGCTGACCACAAATAACGCCTTTAATTACAGGCGCCCCAGATTTTAAAATGCCTGCCCA
                                        ──→ ←──        ──→←──
```

Figure 11: **Alignment of the probes used in the band-shift assays to the upstream region of the NADP-GDH gene.**

The top line represents the upstream region from the GDH gene with the AT-rich and the g-CAB elements indicated by heavy overlines. Tetranucleotide sequences related to g-CAB are indicated by heavy underlines. The opposing arrows represent inverted repeat sequences. The numbering of the nucleotides in the sequence is relative to the zero position assigned to the A of the ATG translation initiation codon. The duplex probes are represented by double lines. See table II for the sequences of the oligonucleotide probes.
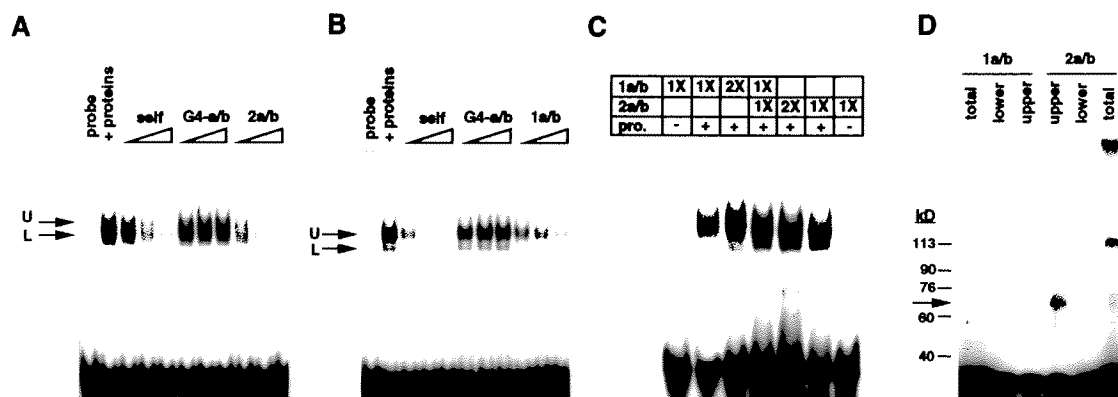
Figure 12: **Binding assays with probes containing either the AT-rich or the g-CAB sequence elements.**
Band-shift assays with either the duplex oligonucleotide 1a/b containing the AT-rich element (**A**) or with the duplex oligonucleotide 2a/b containing the g-CAB element (**B**) as probes. The doublets in the shifted complexes, referred to as the upper (U) and lower (L) bands, observed with each of the two probes, are indicated by the arrows. Increasing amounts of the unlabeled duplex oligonucleotides (20, 40 and 80 fold molar excess relative to the probe) used as competitors are indicated at the top. (**C**) Duplex oligonucleotides 1a/b and 2a/b were used either alone or together as probes, with or without nuclear proteins (pro.) in binding assays as indicated in the table at the top. The amount of probe used in each binding assay was either 0.05 pmoles (1X) or 0.10 pmoles (2X). All other conditions in the binding assays are as described in "Materials and Methods". (**D**) UV cross-linking of proteins bound to duplex oligonucleotides 1a/b and 2a/b was done as described in "Materials and Methods". The autoradiograph from the SDS-polyacrylamide gel is presented. The two outer lanes are the total UV cross-linked binding reactions and the middle lanes are the upper and lower bands in each complex isolated after UV cross-linking and electrophoresis through a standard band-shift gel. The sizes from a protein molecular weight standard are indicated on the left. The arrow indicated the position of the 68 kD protein.

approximately equal affinities. This observation suggested that these two probes are binding the same protein or proteins in the nuclear extract. To examine this phenomenon further, the two duplex probes were incubated together in a band-shift assay (Fig. 12C). No additional shifted complexes were observed in this experiment; this suggests that there is no interaction between the protein binding to these two probes.

The results from the experiments illustrated in figure 12 indicate (i) that the two duplex probes, containing the AT-rich and the g-CAB element respectively, bind to the same protein or proteins in the nuclear extract and (ii) that each probe forms two different complexes with slightly different electrophoretic mobilities. The relationship between these different complexes and the nature of the protein or proteins bound to each probe was analyzed by UV cross-linking. Duplicates of the binding reactions with probes 1a/b and 2a/b were subjected to UV irradiation and one set of each reaction was subjected to electrophoresis through a standard band-shift gel. The upper and lower shifted bands that were observed with each probe were cut out separately from the gel and were loaded on to a SDS-polyacrylamide gel along with the remaining set of UV treated binding reactions. The autoradiograph of this SDS-polyacrylamide gel showed that the upper and lower bands obtained from both oligonucleotide probes migrated at the same molecular weight of approximately 68 kD (Fig. 12D). A band of this molecular weight was also apparent in the unfractionated binding reactions. In a control experiment, gel slices of the shifted complexes were cut from band-shift gels of identical binding reactions that were not cross-linked. DNA recovered from these gel slices was subjected to electrophoresis on a SDS-polyacrylamide gel. No radiolabeled bands other than the bands representing the free probes were observed in the autoradiograph of this gel (data not shown).

The next set of experiments used a 54 bp fragment (GDH PCR54) containing both the AT-rich and the g-CAB motifs (Fig. 11). When this fragment was radiolabeled and

used in band-shift assays, two shifted complexes were observed which I refer to as C1 and C2 (Fig. 13). These complexes represented specific binding of proteins by GDH PCR54 as the formation of these complexes was not inhibited by the addition of an excess of duplex oligonucleotides containing either the yeast GAL4 binding site (G4-a/b) or the adenovirus EIB TATA-box (E1B-a/b). Formation of the complexes was inhibited by the addition of oligonucleotides containing either the AT-rich element (1a/b) or the g-CAB element (2a/b). Furthermore, the two complexes appear to exhibit differential competition, with the upper C2 band being more sensitive to competition than the lower C1 band (Fig. 13A). This differential competition of the two complexes may be related to the ordered appearance of the complexes in the presence of increasing protein concentration in binding reactions with the GDH PCR 54 probe (Fig. 13B). The appearance of the upper complex (C2) requires a higher concentration of protein than that needed for the appearance of the lower complex (C1).

A duplex oligonucleotide (3a/b), representing the region between the g-CAB and the AT-rich motifs (Fig. 11) was radiolabeled and used in band-shift assays to determine its ability to bind proteins (Fig. 14A). Protein binding by this probe cannot be inhibited by duplex oligonucleotides containing the yeast GAL4 binding site (G4-a/b) but is inhibited by oligonucleotides containing the AT-rich (1a/b) and the g-CAB (2a/b) elements. However, the competition for protein binding to the 3a/b probe by an excess of unlabeled 3a/b is weak as compared to the competition by 1a/b and 2a/b. Oligonucleotide 3a/b also acted as a weak competitor against protein binding to oligonucleotides 1a/b, 2a/b and GDH PCR54 (Fig. 14B).

Protein binding to single-stranded DNA

Since I had originally interpreted the above results to indicate that the 68 kD protein binds specifically to double-stranded DNA, I next wanted to determine whether the protein can also bind single-stranded DNA. Single-stranded oligonucleotides were
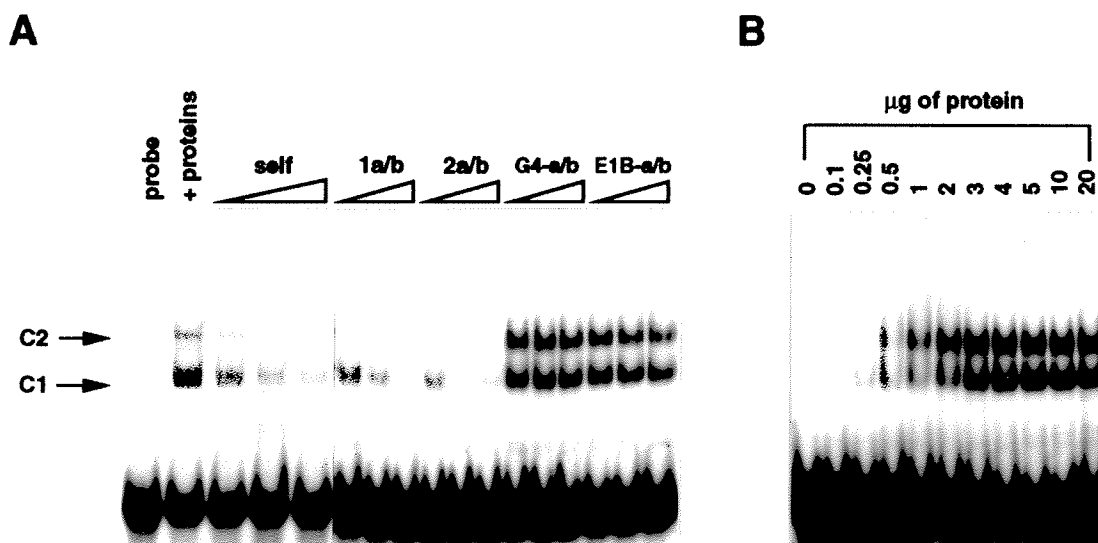
Figure 13: **Band-shift assays with GDH PCR54 containing both the AT-rich and the g-CAB elements.**
(A) Competitors used against protein binding to the GDH PCR54 probe are indicated at the top. Competition was at 20, 40 and 80 fold molar excess of the unlabeled DNAs relative to the probe. The two shifted complexes, C1 and C2, are indicated by the arrows. (B) Increasing amounts of protein extract was added to binding reactions with the GDH PCR54 probe. The amount of NaCl added to each of the binding reactions was adjusted to maintain a constant concentration of 100 mM to compensate for the increasing amounts of NaCl introduced with the increasing amounts of the protein extracts added.
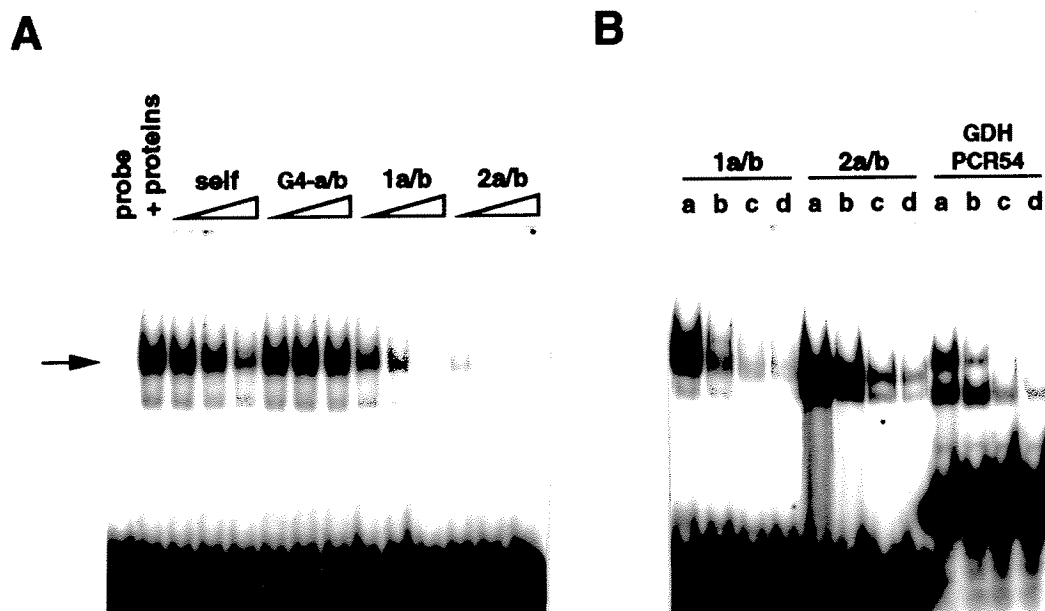
Figure 14: **Band-shift assays with duplex oligonucleotide 3a/b containing neither the AT-rich nor the g-CAB elements.**
(**A**) Competitors used against protein binding to the duplex 3a/b probe are indicated at the top. Competition was at 20, 40, and 80 fold molar excess of the unlabeled oligonucleotides relative to the probe. The shifted complex representing specific binding is indicated by the arrow. (**B**) Duplex oligonucleotide 3a/b was used as a competitor in protein binding assays against the duplex probes listed at the top. Lanes: a, probe alone; b, probe plus proteins; c and d, reactions with 40 and 80 fold molar excess of unlabeled duplex 3A/B as competitor, respectively.

used as either probes or competitors in band-shift assays (Fig. 15). Single-stranded oligonucleotides (both (+) and (-) strands) containing either the AT-rich (1a, 1b) or the g-CAB (2a, 2b) element can compete for protein binding to GDH PCR54 fragment (Fig. 15A). In contrast, only the (+) strand of the region between the AT-rich and the g-CAB element (oligonucleotide 3a) can compete for binding. Neither strand of the DNAs containing the yeast GAL4 binding site (G4-a, G4-b) or the adenovirus E1B TATA box (E1-a, E1-b) can act as competitor for protein binding to GDH PCR54.

When the single-stranded oligonucleotides were used as probes in band-shift assays, both (+) and (-) strands containing the AT-rich (1a, 1b) and the g-CAB (2a, 2b) elements bound proteins whereas only the (+) strand of the region between the two elements (3a) bound protein. The protein binding to these single-stranded probes can be inhibited by the addition of an excess of unlabeled GDH PCR54 (Fig. 15B).

DNA sequence requirements for protein binding

Examination of the oligonucleotides used in the binding assays revealed that there is a TTT sequence in all the single-stranded DNA probes that bind protein (Table II). This TTT sequence is absent in the probes that did not bind. To investigate the possibility that a TTT sequence is required for the binding of the 68 kD protein, oligonucleotides containing point mutations were tested in binding assays (Fig. 15B). Probe 3a*, which differs from oligonucleotide 3a by a single point mutation that converts a TTT trinucleotide to a TGT sequence, was unable to bind the 68 kD protein. Probe 3b*, which differ from oligonucleotide 3b by a TGT to TTT conversion, did bind the 68 kD protein.

Similarly, complementary oligonucleotides 4b and 4a, representing a sequence of the upstream region of the GDH gene containing one and no TTT sequence, respectively (Fig. 11), can bind the 68 kD protein when annealed together as duplex DNA but only oligonucleotide 4a can bind as a single-stranded DNA probe (Fig. 15C). Furthermore,
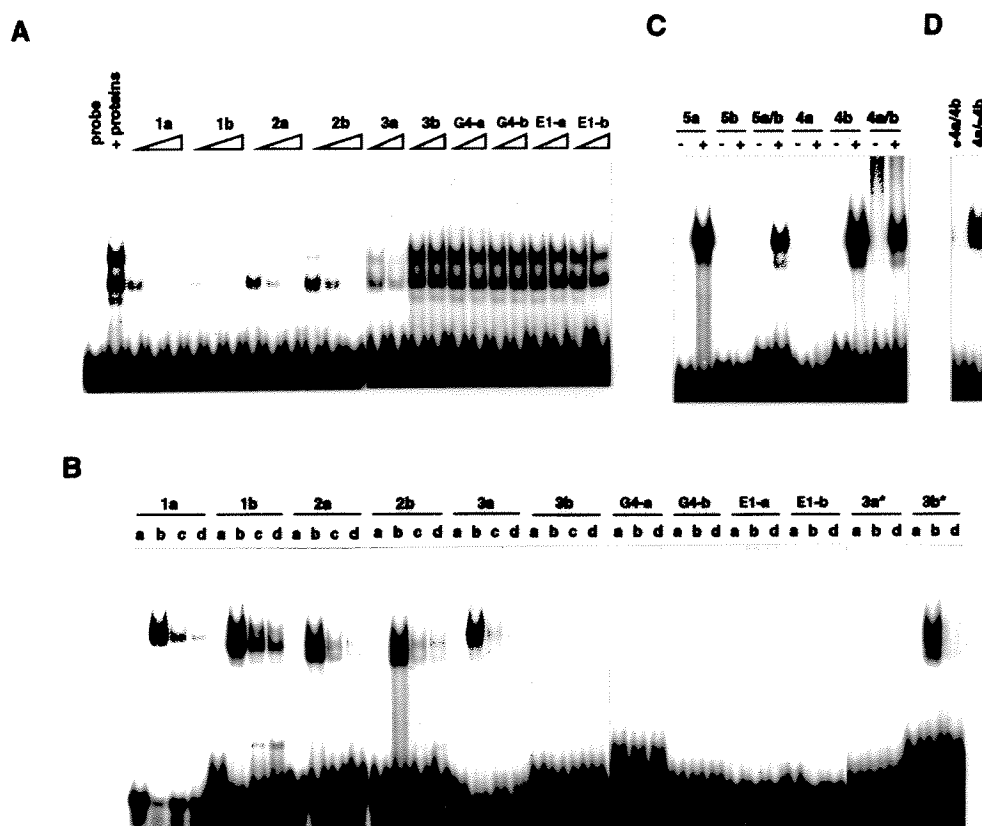
Figure 15: **Band-shift assays with single-stranded oligonucleotides as either probes or competitors.**
Oligonucleotides representing the (+) DNA strand are given the 'a' designation and those representing the (-) strand are given the 'b' designation in the name. (A) Protein binding to the duplex GDH PCR54 probe was inhibited by increasing amounts of the single-stranded oligonucleotides listed at the top. For competitors 1a, 1b, 2a, and 2b, the amounts used were 20, 40 and 80 fold molar excess over the probe. For 3a, 3b, G4-a, G4-b, E1-a, and E1-b, 40 and 80 fold molar excess over the probe were used. (B) The different single-stranded probes used in the binding assays are listed at the top. The probes 3a* and 3b* are oligonucleotides containing point mutations (see main text and table I). Lanes: a, probe alone; b, probe plus proteins; c and d, reactions with 40 and 80 fold molar excess of unlabeled duplex GDH PCR54 as competitor, respectively. (C) The different probes used in the binding assays are listed at the top. Lanes: (-) probe alone; (+) probe plus proteins. (D) •4a/4b represents the binding reaction with the duplex probe consisting of radiolabeled oligonucleotide 4a annealed to unlabeled oligonucleotide 4b. Similarly, 4a/•4b represents the binding reaction with unlabeled 4a annealed to radiolabeled 4b.

binding of the 68 kD protein to 4a can be inhibited by the addition of oligo(dT)$_{12-18}$ (data not shown). Although these results suggest that a minimum sequence of a TTT trinucleotide is required for the binding of the 68 kD protein, I wanted to determine if there are other sequence requirements. I tested the binding ability of the 68 kD protein to a pair of complementary oligonucleotides (5A and 5B) derived from within the coding region of the GDH gene (Fig. 15C). These oligonucleotides represents a portion of the GDH gene encoding amino acids 103 to 109 (LKFLGFE) at nucleotide position 531-550 in the sequence presented in figure 5. Other than a single trinucleotide TTT sequence in 5A, there is no sequence similarity between these oligonucleotides and those tested previously. Since the single-stranded oligonucleotide 5A and duplex oligonucleotide 5a/b can also bind the 68 kD protein, I must conclude that a single TTT sequence is sufficient for binding of this protein to both single and double-stranded DNA. Table II presents a summary of the band-shift assay results.

Although I assumed that the complexes observed with the duplex oligonucleotide probes represented binding of the 68 kD protein to the double-stranded DNA, it is possible that the protein is binding to single-stranded DNA melted from the duplexes. To investigate this possibility, oligonucleotide 4a, which does not contain any TTT sequences, was radiolabeled and annealed to its unlabeled complementary oligonucleotide 4b, which contains one TTT sequence. After annealing, this mixture was subjected to electrophoresis through a 20% non-denaturing polyacrylamide gel beside a sample of single-stranded radiolabeled 4a. Examination of the autoradiograph obtained from this gel revealed that >90% of the radiolabeled oligonucleotide 4a was in the duplex form (data not shown). When this mixture was used in the band-shift assay, no shifted complex was observed. In the complementary experiment with the duplex probe containing unlabeled oligonucleotide 4a and radiolabeled oligonucleotide 4b, a shifted complex was observed (Fig. 15D). Thus, only single-stranded DNA binding by the 68 kD protein was detected in this experiment.

Table II: List of oligonucleotides used in band-shift assays

single-strand name[1]	sequence[2] (5'→3')	double-strand (duplex) name	mobility shift
1a	CCCAGA<u>TTTT</u>AAAATGCCTG	1a/b	yes
1b	GGGCAGGCA<u>TTTT</u>AAAATCTGGG		yes
2a	GACCACAAATAACGCC<u>TTT</u>	2a/b	yes
2b	TTAAAGGCGTTA<u>TTT</u>GTGGTC		yes
3a	CGCC<u>TTT</u>AATTACAGGCGC	3a/b	yes
3b	GGGGCGCCTGTAATTAAAGGCG		no
3a*	CGCCTGTAATTACAGGCGC		no
3b*	GGGGCGCC<u>TTT</u>AATTAAAGGCG		yes
4a	AAGCTGACCACAAATAACG	4a/b	no
4b	GGCGTTA<u>TTT</u>GTGGTCAGCTT		yes
5a	TCAAGTTCCTCGG<u>TTT</u>CGAG	5a/b	yes
5b	TGCTCGAAACCGAGGAACTTGA		no
G4-a	TCGACGGAGTACTGTCCTCCG	G4-a/b	no
G4-b	TCGACGGAGGACAGTACTCCG		no
E1B-a	CTAGAGGGTATATAATGCGCCAGCT	E1B-a/b	no
E1B-b	GGCGCATTATATACCCT		no

[1]Oligonucleotides representing the (+) DNA strand are given the 'a' designation and those representing the (-) strand are given the 'b' designation in the name.

[2]Sequences with three or more T's in a row are underlined and (●) indicates the nucleotide which was altered in the mutant oligonucleotides.

Protein binding to RNA

The ability of the 68 kD protein to bind to RNA was examined. Transcripts derived from the 5' flanking region of the GDH gene (position -223 to +21 relative to the zero position assigned to the A of the ATG translation initiation codon; nucleotide position 1 to 245 in the sequence presented in figure 5) were synthesized *in vitro* using an SP6 promoter system. The transcribed RNA contained three separate UUU sequences corresponding to three TTT sequences on the (+) DNA strand. When the radiolabeled RNA was incubated with proteins from the nuclear extract, complexes were obtained but these could not be inhibited by an excess of unlabeled oligo(dT)$_{12-18}$ or by oligonucleotide 4b (data not shown). Similarly, an excess of unlabeled RNA could not inhibit protein binding to radiolabeled oligonucleotide 4b (data not shown). These results suggest that the protein which is binding to TTT sequences on single-stranded DNA does not bind RNA containing UUU sequences.

Discussion

Distribution of the AT-rich and the g-CAB elements

Two conserved sequences, referred to as the AT-rich and g-CAB elements, have been identified in the region upstream of the GDH gene and in other *Giardia* genes (Fig. 8). The AT-rich element has the consensus A(A/T)TTAAAAATG and contains both the start sites for transcription initiation and the ATG translation initiation codon. The g-CAB element with the consensus CAAAT, is present in one or more copies between 25 to 50 nucleotides upstream of the sites of transcription initiation. Both of these elements have a high AT content, in contrast to the normally GC-rich content of the *Giardia* genome. Melting of DNA at sites within these elements may facilitate the assembly of a RNA polymerase II preinitiation complex. In addition, the upstream region of the GDH gene contains two inverted repeats: one 5 bp long and encompassing the AT-rich element and the other 9 bp long with one mismatch (Fig. 9). Inverted repeat sequences are common recognition motifs for a number of regulatory proteins including transcription factors. Furthermore, these AT-rich and g-CAB elements are found in approximately the same position in the upstream region of all *Giardia* genes characterized, with the exception of the genes encoding variant surface antigens (Fig. 8).

The absence of one or both of these elements from the variant-specific surface antigen genes (VSP 1267 and TSA 417) may reflect a different mechanism for the regulation of these genes. This is not entirely unexpected since the expression of the antigen genes is involved in the phenomenon of antigenic variation, a process that might require more complex regulation. Another difference between these genes is that the transcripts from the antigen genes contain an untranslated RNA leader sequence consisting of just a single nucleotide whereas the transcripts from the other genes contain leaders that are three to six nucleotides in length.

In the parasite *Trypanosoma brucei*, in which antigenic variation also occurs, the antigen genes are suspected to be transcribed by RNA polymerase I, as their transcription is α-amanitin resistant (Rudenko *et al.*, 1989; Zomerdijk *et al.*, 1991). This is in contrast to the α-amanitin sensitive transcription of other protein-encoding genes. Thus, the A/T rich and the g-CAB elements may only be prevalent in a class of *Giardia* genes that excludes the variant antigen genes because of differences in either regulation requirements or transcription by different RNA polymerases, or a combination of both.

Sequence specificity of POT binding

A 68 kD protein that can bind in a sequence specific manner to single-stranded DNA from the upstream region of the GDH gene was identified from a *Giardia* nuclear extract. Sequence comparison of probes that are recognized by this 68 kD protein and binding studies on mutant probes indicated that this protein requires a minimum recognition sequence of TTT. This protein is named POT for its poly(T) binding ability.

In the upstream region of the GDH gene there are four potential binding sites for POT; two TTT sequences are on the (+) strand and two are on the (-) strand. Of these, two are within the AT-rich element, one on each strand, and one is within the g-CAB element on the (-) strand. The fourth site is located on the (+) strand between the two elements. The locations of these binding sites are consistent with the observation that oligonucleotide probes representing sub-sections of this region can cross-compete with each other for the binding of POT. Duplex oligonucleotides 1a/b and 2a/b contain two potential binding sites, one on each DNA strand, and 3a/b contains only a single binding site on its (+) strand. These differences in the numbers of binding sites are consistent with the observation that oligonucleotides 1a/b and 2a/b have approximately equal affinities for POT whereas 3a/b has a lower affinity in band-shift competition assays. When the different single-stranded oligonucleotides (1a, 1b, 2a, 2b, 3a, 3b*, 4a, 5a), each containing a single potential POT binding site, were used in cross-

competition binding assays, they all appear to have approximately equal affinities for POT (data not shown).

POT binding sites in the GDH PCR54 fragments

Although there are four potential binding sites for POT in the GDH PCR54 fragment, only two complexes (C1 and C2) are observed in gel mobility shift assays (Fig. 11). The differential competition and dependency on protein concentration for the formation of C1 and C2 suggest that these complexes represent two levels of POT binding. If POT binds as a monomer, then C1 and C2 may represent binding at one and two sites, respectively. Steric hindrance could prevent the binding of POT to more than two of the four potential sites at any one time. Alternatively, if POT is binding as a dimer, then C1 may represent binding of one dimer to two of the four sites and C2 may represent binding of two dimers, one subunit binding at each of the four sites. In the above explanations, I assumed that the GDH PCR54 probe remained as a duplex with localized melting at the TTT containing sequences where POT binds. If the GDH PCR54 is completely melted in the complexes, then C1 may represent binding of one of the two sites available on each DNA strand and C2 may represent binding at both sites.

Doublet band formations in some POT-DNA complexes

The binding of POT to some DNA probes resulted in shifted complexes consisting of doublet bands. These doublets are distinct from the C1 and C2 complexes observed with the GDH PCR54 probe by three criteria. First, the doublet bands are not subjected to differential competition. Second, their appearance is simultaneous with increasing protein concentrations and third, they cannot be correlated with the number of TTT sites on the probe. Finally, the doublets appear with some but not all single and double-stranded probes. For example, doublets were observed with probes 1a/b, 2a/b, 1b, 2b,

3b* and 4b but not with probes 3a/b, 5a/b, 1a, 3a and 5a. Thus, these bands are unlikely to represent different levels of protein binding.

The doublet bands formed with duplex oligonucleotides 1a/b and 2a/b were analyzed separately by UV cross-linking (Fig. 10D). Both components of the doublets exhibit cross-links to a protein of 68 kD. Since it was possible to separate the two bands of the doublet on a non-denaturing polyacrylamide gel (used in the band-shift analysis) but not on a denaturing gel, I can speculate about some possible explanations for these observations.

One possibility is that the two bands represent DNA binding by two different proteins with the same molecular weights but with different conformations under native gel conditions. The probe specific appearance of the doublet formation may indicate that at least one of the proteins has a more restricted recognition site.

Alternatively, the two bands may represent DNA binding by the same protein but with modifications affecting the net charge of the protein. One such modification could be either a phosphorylation or a dephosphorylation so that the two protein forms would be separated under native gel conditions due to the differences in net charges but would not be separated under denaturing gel conditions due to the slight change in the sizes of the two forms of the protein.

A third possibility is that protein binding to some DNA sequences may induce DNA bending which results in two protein-DNA isomers with slightly different gel mobilities. If this conformational change is dependent on the nucleotide composition of both the protein binding site and the flanking sequences, then doublet formation would only be observed upon protein binding to a subset of probes. The binding of the products of the oncogenes Jun and Fos to the AP-1 site is an example of proteins inducing a conformational change to the DNA. Binding of Fos-Jun heterodimers bend DNA toward the major groove whereas Jun homodimers bend DNA toward the minor

groove and these species can be distinguished by gel mobility assays (Kerppola and Curran, 1991). I am unable to distinguish between these alternative from our results.

Roles of POT in gene regulation

Regulatory proteins that can interact with both single and double-stranded DNA in a sequence specific manner to control either replication or transcription have been identified in mammalian muscle cells, HeLa cells, and yeast (Hofmann and Gasser, 1991; Santoro et al., 1991; Bergemann and Johnson, 1992). It is clear from the results that POT can bind to TTT sequences on single-stranded DNA but it is uncertain whether POT can also bind to double-stranded DNA. Although POT binding is observed in band-shift assays with double-stranded oligonucleotide probes, it is possible that POT is binding to DNA that remains single-stranded after annealing of the complementary oligonucleotides. This is less likely to be the explanation in the experiments with the double-stranded GDH PCR54 fragment which was generated by PCR and was gel purified away from any possible contaminating single-stranded DNA prior to its use. Since GDH PCR54 is both a substrate and a competitor for POT binding, a helicase activity in the nuclear extract may be inferred. However, I do not know if this helicase activity is associated with POT or with another protein in the nuclear extract.

Results from band-shift assays with the 4a/b duplexes labeled on only one strand suggest that short oligonucleotide probes are completely denatured when bound to POT. However, I do not know if GDH PCR54 is completely denatured by this helicase activity or whether there is only localized melting. Nevertheless, the ability of POT to bind to single-stranded DNA may play an important role in Giardia gene regulation since the melting of duplex DNA and its maintenance in a single-stranded form is an essential feature in both transcription and DNA replication in other eukaryotes. The binding of POT to the TTT sequences found in the AT-rich elements, the g-CAB elements, and to other TTT sequences in the upstream region of the GDH gene may

facilitate the melting of the DNA and allow the formation of a transcription initiation complex to occur within this region.

Is POT a homologue of a transcription factor previously characterized in higher eukaryotes? One possibility is that POT is related to the TATA-binding protein (TBP) and the AT-rich sequence is functionally equivalent to the TATA-box. Another possibility is that the AT-rich motif functions as an initiator element and *Giardia* genes have TATA-less promoters. In this scenario, POT binds to the initiator and recruits a TBP-like protein to this putative promoter region in an analogous manner as the TBP-tethering factors in others eukaryotes. A final alternative is that POT is unrelated to any known transcription factor and *Giardia* utilizes an unique system for the control of transcription initiation.

To my knowledge, this is the first example of a DNA binding protein with a sequence specificity of only three nucleotides. Our results suggest that POT has the potential to bind to every site in the *Giardia* genome where there are at least three T's in a row. Although the protein encoding regions in the *Giardia* genome have a strong GC bias, TTT sequences do occasionally occur in these and other non-promoter regions. What would be the function of POT if it binds to all these sites? One alternative possibility is that POT is not directly involved with transcription but may be associated with other functions such as recombination or chromatin organization. The problem with this alternative is that these types of proteins, for example topoisomerases and histones, normally do not have a specific sequence requirement for their binding to DNA. Another possibility is that I am overlooking other physical requirements in our *in vitro* binding assays that limit the binding of POT to a subset of TTT sequences in the *Giardia* genome *in vivo*. Physical requirements could include chromatin structure and DNA methylation. Yet another possibility is that POT binds cooperatively as multimers to a number of TTT sites which are located close together as in the upstream putative promoter region of the GDH gene and not to relatively isolated TTT sites as found

within non-promoter regions. A final alternative is that the binding or activity of POT is mediated by other protein factors that I was not able to detect in these binding assays.

In summary, two conserved sequence motifs, the AT-rich and the g-CAB elements, are found in front of a variety of *Giardia* genes. Their conservation and location relative to the sites of transcription initiation suggest that these elements may be involved in the regulation of transcription. A 68 kD protein, called POT, binds to the TTT trinucleotides found on either one or both strands of the DNA encoding these elements in the upstream region of the GDH gene. Because of a current lack of a transformation system or an *in vitro* transcription system for *Giardia*, it is not yet possible to define the precise function of these elements and the role of POT in the regulation of gene expression in this organism. However, it may be possible to gain a better idea about the function of POT when a more purified preparation of POT is available and when the gene encoding POT is obtained.

Cavalier-Smith (1987) first suggested that *Giardia* has retained characteristics of a primitive eukaryote from the observation that mitochondria and other organelles are absent in this organism. This idea was supported later by phylogenetic analysis of *Giardia's* 16S-like rRNA gene (Sogin et al., 1988). *Giardia* is, therefore, a valuable resource for studying aspects of cellular function in an early eukaryote. The identification of a nuclear protein which can bind specifically to conserved sequence elements in *Giardia* gene promoters may be the first step in understanding the regulation of transcription, and ultimately the control of gene expression in this organism.

References

Adam, R.D. (1991). The biology of *Giardia* spp. *Microbiol. Rev.*, **55:** 706-732.

Adam, R.D., A. Aggarwal, A.A. Lal, V.F. de la Cruz, T. McCutchan, and T.E. Nash (1988). Antigenic variation of a cysteine-rich protein in *Giardia lamblia*. *J. Exp. Med.*, **167:** 109-118.

Adam, R.D., T.E. Nash, and T.E. Wellems (1988). The *Giardia lamblia* trophozoite contains sets of closely related chromosomes. *Nucleic Acids Res.*, **16:** 4555-4567.

Adam, R.D., Y.M. Yang, and T.E. Nash (1992). The cysteine-rich protein gene family of *Giardia lamblia*: loss of the *CRP170* gene in an antigenic variant. *Mol. Cell. Biol.*, **12:** 1194-1201.

Alonso, R.A. and D.A. Peattie (1992). Nucleotide sequence of a second alpha giardin gene and molecular analysis of the alpha giardin genes and transcripts in *Giardia lamblia*. *Mol. Biochem. Parasitol.*, **50:** 95-104.

Andrews, N.C. and D.V. Faller (1991). A rapid micropreparation technique for extraction of DNA-binding proteins from limiting numbers of mammalian cells. *Nucleic Acids Res.*, **19:** 2499.

Benachenhou-Lahfa, N., P. Forterre, and B. Labedan (1993). Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. *J. Mol. Evol.*, **36:** 335-346.

Boothroyd, J.C., A. Wang, D.A. Campbell, and C.C. Wang (1987). An unusually compact ribosomal DNA repeat in the protozoan *Giardia lamblia*. *Nucleic Acids Res.*, **15:** 4065-4084.

Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.*, **72:** 248-254.

Buratowski, S., S. Hahn, L. Guarente, and P.A. Sharp (1989). Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell*, **56:** 549-561.

Cavalier-Smith, T. (1987). Eukaryotes with no mitochondria. *Nature*, **326:** 332-333.

Cazzulo, J.J., C. Nowicki, J.A. Santome, C. Wernstedt, and U. Hellman (1988). Amino acid composition and N-terminal sequence of the NADP-linked glutamate dehydrogenase from *Trypanosoma cruzi*. *FEMS Microbiology Letters*, **56:** 215-220.

Cormack, B.P. and K. Struhl (1992). The TATA-binding protein is required for transcription by all three nuclear RNA polymerases in yeast cells. *Cell,* **69:** 685-696.

Crossley, R. and D.V. Holberton (1983). Characterization of proteins from the cytoskeleton of *Giardia lamblia. J. Cell Sci.,* **59:** 81-103.

Crossley, R. and D.V. Holberton (1985). Assembly of 2.5 nm filaments from giardin, a protein associated with cytoskeletal microtubules in *Giardia. J. Cell Sci.,* **78:** 205-231.

Crossley, R., J. Marshall, J.T. Clark, and D.V. Holberton (1986). Immunocytochemical differentiation of microtubules in the cytoskeleton of *Giardia lamblia* using monoclonal antibodies to alpha-tubulin and polyclonal antibodies to associated low molecular weight proteins. *J. Cell Sci.,* **80:** 233-252.

Dynan, W.S., S. Sazer, R. Tjian, and R.T. Schimke (1986). Transcription factor Sp1 recognizes a DNA sequence in the mouse dihydrofolate reductase promoter. *Nature,* **319:** 246-248.

Dynan, W.S. and R. Tjian (1983). The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell,* **35:** 79-87.

Dynlacht, B.D., T. Hoey, and R. Tjian (1991). Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. *Cell,* **66:** 563-576.

Fan, J.B., S.H. Korman, C.R. Cantor, and C.L. Smith (1991). *Giardia lamblia:* haploid genome size determined by pulsed field gel electrophoresis is less than 12 Mb. *Nucleic Acids Res.,* **19:** 1905-1908.

Favaloro, J., R. Treisman, and R. Kamen (1980). Transcription maps of polyoma virus-specific RNA: analysis by two-dimensional nuclease S1 gel mapping. *Methods Enzymol.,* **65:** 718-749.

Feely, D.E., J.V. Schollmeyer, and S.L. Erlandsen (1982). *Giardia:* distribution of contractile proteins in the attachment organelle. *Exp. Parasitol,* **53:** 145-154.

Feinberg, A. and B. Vogelstein (1982). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.,* **132:** 6-13.

Frohman, M.A., M.K. Dush, and G.R. Martin (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. U.S.A.,* **85:** 8998-9002.

Gillin, F.D., P. Hagblom, J. Harwood, S.B. Aley, D.S. Reiner, M. McCaffery, M. So, and D.G. Guiney (1990). Isolation and expression of the gene for a major surface protein of *Giardia lamblia. Proc. Natl. Acad. Sci. U.S.A.,* **87:** 4463-4467.

Green, M.R. (1992). Transcriptional transgressions. *Nature*, **357:** 364-365.

Hawkins, A.R., S.J. Gurr, P. Montague, and J.R. Kinghorn (1989). Nucleotide sequence and regulation of expression of the *Aspergillus nidulans gdh*A gene encoding NADP dependent glutamate dehydrogenase. *Mol. Gen. Genet.*, **218:** 105-111.

Healey, A., R. Mitchell, J.A. Upcroft, P.F.L. Boreham, and P. Upcroft (1990). Complete nucleotide sequence of the ribosomal RNA tandem repeat unit from *Giardia intestinalis*. *Nucleic Acids Res.*, **18:** 4006.

Helmann, J.D. and M.J. Chamberlin (1988). Structure and function of bacterial sigma factors. *Ann. Rev. Biochem.*, **57:** 839-872.

Henikoff, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28:** 351-359.

Hoffman, A., E. Sinn, T. Yamamoto, J. Wang, A. Roy, M. Horikoshi, and R.G. Roeder (1990). Highly conserved core domain and unique N terminus with presumptive regulatory motifs in a human TATA factor (TFIID). *Nature*, **346:** 387-390.

Holberton, D., D.A. Baker, and J. Marshall (1988). Segmented α-helical coiled-coil structure of the protein giardin from the *Giardia* cytoskeleton. *J. Mol. Biol.*, **204:** 789-795.

Holberton, D.V. (1981). Arrangement of subunits in microribbons from *Giardia*. *J. Cell Sci.*, **47:** 167-185.

Horikoshi, M., C.K. Wang, H. Fujii, J.A. Cromlish, P.A. Weil, and R.G. Roeder (1989). Cloning and structure of a yeast gene encoding a general transcription initiation factor TFIID that binds to the TATA box. *Nature*, **341:** 299-303.

Huang, J. and L.H.T. Van der Ploeg (1991). Maturation of polycistronic pre-mRNA in *Trypanosoma brucei*: analysis of *trans* splicing and poly(A) addition at nascent RNA transcripts from the *hsp70* locus. *Mol. Cell. Biol.*, **11:** 3180-3190.

Jarroll, E.L., P. Manning, A. Berrada, D. Hare, and D.G. Lindmark (1989). Biochemistry and metabolism of *Giardia lamblia*. *J. Protozool.*, **36:** 190-197.

Jarroll, E.L., P. Manning, D.G. Lindmark, J.R. Coggins, and S.L. Erlandsen (1989). *Giardia* cyst wall-specific carbohydrate: evidence for the presence of galactosamine. *Mol. Biochem. Parasitol.*, **32:** 121-132.

Johnston, M. and R.W. Davis (1984). Sequences that regulate the divergent GAL1-GAL10 promoter in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **4:** 1440-1448.

Kabnick, K.S. and D.A. Peattie (1990). *In situ* analyses reveal that the two nuclei of *Giardia lamblia* are equivalent. *J. Cell Sci.*, **95**: 353-360.

Keister, D.B. (1983). Axenic culture of *Giardia lamblia* in TYI-S-33 medium supplemented with bile. *Trans. R. Soc. Trop. Med. Hyg.*, **77**: 487-488.

Kerppola, T.K. and T. Curran (1991). Fos-Jun heterodimers and Jun homodimers bend DNA in opposite orientations: implications for transcription factor cooperativity. *Cell*, **66**: 317-326.

Kirk-Mason, K.E., M.J. Turner, and P.R. Chakraborty (1989). Evidence for unusually short tubulin mRNA leaders and characterization of tubulin genes in *Giardia lamblia*. *Mol. Biochem. Parasitol.*, **36**: 87-100.

Kozak, M. (1989). The scanning model for translation: an update. *J. Cell Biol.*, **108**: 229-241.

La Thangue, N.B. and P.W.J. Rigby (1988). *Trans*-acting protein factors and the regulation of eukaryotic transcription, pp. 1-41. In Transcription and Splicing, Frontiers in Molecular Biology, B.D. Hames and D.M. Glover (Eds.). IRL Press, Oxford, England, and Washington, D.C.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of the bacteriophage T4. *Nature*, **227**: 680-685.

Lindmark, D.G. (1980). Energy metabolism of the anaerobic protozoon, *Giardia lamblia*. *Mol. Biochem. Parasitol.*, **1**: 1-12.

Lindmark, D.G. (1988). *Giardia lamblia*: localization of hydrolase activities in lysosome-like organelles of trophozoites. *Exp. Parasitol.*, **65**: 141-147.

Mattaj, I.W., M.J. McPherson, and J.C. Wootton (1982). Localisation of a strongly conserved section of coding sequence in glutamate dehydrogenase genes. *FEBS Letters*, **147**: 21-25.

Maxam, A.M. and W. Gilbert (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.*, **65**: 499-560.

Meyer, E.A. (1976). *Giardia lamblia*: isolation and axenic cultivation. *Exp. Parasitol.*, **39**: 101-105.

Miller, S.M. and B. Magasanik (1990). Role of NAD-linked glutamate dehydrogenase in nitrogen metabolism in *Saccharomyces cerevisiae*. *J. Bacteriol.*, **172**: 4927-4935.

Mowatt, M., A. Aggarwal, and T. Nash (1992). Carboxy-terminal sequence conversion among variant-specific surface proteins of *Giardia lamblia*. *Mol. Biochem. Parasitol.*, **49:** 215-227.

Muller, M. (1988). Energy metabolism of protozoa without mitochondria. *Ann. Rev. Microbiol.*, **42:** 465-488.

Murtagh, J.J., Jr., M.R. Mowatt, C.-M. Lee, F.-J.S. Lee, K. Mishima, T.E. Nash, J. Moss, and M. Vaughan (1992). Guanine nucleotide-binding proteins in the intestinal parasite *Giardia lamblia*. *J. Biol. Chem.*, **267:** 9654-9662.

Nakatani, Y., M. Horikoshi, M. Brenner, T. Yamamoto, F. Besnard, R.G. Roeder, and E. Freese (1990). A downstream initiation element required for efficient TATA box binding and *in vitro* function of TFIID. *Nature*, **348:** 86-88.

Nash, T. (1992). Surface antigen variability and variation in *Giardia lamblia*. *Parasitol. Today*, **8:** 229-234.

Nash, T.E., S.M. Banks, D.W. Alling, J.W. Merrit Jr., and J.T. Conrad (1990). Frequency of variant antigens in *Giardia lamblia*. *Exp. Parasitol.*, **71:** 415-421.

Nash, T.E., J.W.M. Jr., and J.T. Conrad (1991). Isolate and epitope variability in susceptibility of *Giardia lamblia* to intestinal proteases. *Infect. Immun.*, **59:** 1334-1340.

Nash, T.E. and M.R. Mowatt (1992). Characterization of a *Giardia lamblia* variant-specific protein (VSP) gene from isolate GS/M and estimation of the VSP gene repertoire size. *Mol. Biochem. Parasitol.*, **51:** 219-228.

Nash, T.E. and M.R. Mowatt (1993). Variant-specific surface proteins of *Giardia lamblia* are zinc-binding proteins. *Proc. Natl. Acad. Sci. U.S.A.*, In press.

Newman, A. (1987). Specific accessory sequences in *Saccharomyces cerevisiae* introns control assembly of pre-mRNA into spliceosomes. *EMBO J.*, **6:** 3833-3839.

Newton, C.H., J. Yee, L.C. Shimmin, and P.P. Dennis (1990). A family of genes encode the multiple forms of the yeast L12 equivalent proteins and a single form of the L10 equivalent proteins. *J. Bacteriol.*, **172:** 579-588.

Paget, T.A., M.L. Kelly, E.L. Jarroll, D.G. Lindmark, and D. Lloyd (1993). The effects of oxygen on fermentation in *Giardia lamblia*. *Mol. Biochem. Parasitol.*, **57:** 65-72.

Peattie, D.A., R.A. Alonso, A. Hein, and J.P. Caulfield (1989). Ultrastructural localization of giardins to the edges of disk microribbons of *Giardia lamblia* and the nucleotide and deduced protein sequence of alpha giardin. *J. Cell Biol.*, **109:** 2323-2335.

Peterson, M.G., N. Tanese, B.F. Pugh, and R. Tjian (1990). Functional domains and upstream activation properties of cloned human TATA binding protein. *Science*, **248:** 1625-1630.

Pugh, B.F. and R. Tjian (1990). Mechanism of transcriptional activation by Sp1: evidence for coactivators. *Cell*, **61:** 1187-1197.

Pugh, B.F. and R. Tjian (1991). Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes & Dev.*, **5:** 1935-1945.

Reiner, D.S., M. McCaffery, and F.D. Gillin (1990). Sorting of cyst wall proteins to a regulated secretory pathway during differentiation of the primitive eukaryote, *Giardia lamblia. Eur. J. Cell Biol.*, **53:** 142-153.

Reznikoff, W.S., D.A. Siegele, D.W. Cowing, and C.A. Gross (1985). The regulation of transcription initiation in bacteria. *Ann. Rev. Genet.*, **19:** 355-387.

Rigby, P.W.J. (1993). Three in one and one in three: it all depends on TBP. *Cell*, **72:** 7-10.

Roeder, R.G. (1991). The complexities of eukaryotic transcription initiation: regulation of preinitiation complex assembly. *Trends Bio. Sci.*, **16:** 402-408.

Roy, A.L., M. Meisterernst, P. Pognonec, and R.G. Roeder (1991). Cooperative interaction of an initiator-binding transcription initiation factor and the helix-loop-helix activator USF. *Nature*, **354:** 245-248.

Rudenko, G., D. Bishop, K. Gottesdiener, and L.H. Van der Ploeg (1989). alpha-amanitin resistant transcription of protein coding genes in insect and bloodstream form *Trypanosoma brucei. EMBO J.*, **8:** 4259-4263.

Sanger, F., S. Nicklen, and A.R. Coulson (1977). DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, **74:** 5463-5467.

Santoro, I.M., T.-M. Yi, and K. Walsh (1991). Identification of single-stranded-DNA-binding proteins that interact with muscle gene elements. *Mol. Cell. Biol.*, **11:** 1944-1953.

Sehgal, A., N. Patil, and M. Chao (1988). A constitutive promoter directs expression of the nerve growth factor receptor gene. *Mol. Cell. Biol.*, **8:** 3160-3167.

Seto, E., Y. Shi, and T. Shenk (1991). YY1 is an initiator sequence-binding protein that directs and activates transcription *in vitro. Nature*, **354:** 241-245.

Sharp, P.A. (1992). TATA-binding protein is a classless factor. *Cell*, **68:** 819-821.

Sheffield, H.G. and B. Bjorvatn (1977). Ultrastructure of the cyst of *Giardia lamblia*. *Am. J. Trop. Med. Hyg.*, **26:** 23-30.

Shi, Y., E. Seto, L.-S. Chang, and T. Shenk (1991). Transcription repression by YY1, a human GLI-Kruppel-related protein, and relief of repression by adenovirus E1A protein. *Cell*, **67:** 377-388.

Smale, S.T. and D. Baltimore (1989). The "initiator" as a transcription control element. *Cell*, **57:** 103-113.

Smale, S.T., M.C. Schmidt, A.J. Berk, and D. Baltimore (1990). Transcriptional activation by Sp1 as directed through TATA or initiator: Specific requirement for mammalian transcription factor IID. *Proc. Natl. Acad. Sci. U.S.A.*, **87:** 4509-4513.

Smith, E.L., B.M. Austen, K.M. Blumenthal, and J.F. Nyc (1975). Glutamate dehydrogenases, pp. 293-367. In The enzymes, Vol. XI, P. D. Boyer (Eds.), New York, Academic Press.

Sogin, M.L., J.H. Gunderson, H.J. Elwood, R.A. Alonso, and D.A. Peattie (1989). Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. *Science*, **243:** 75-77.

Stevens, L., D. Duncan, and P. Robertson (1989). Purification and characterisation of NAD-glutamate dehydrogenase from *Aspergillus nidulans*. *FEMS Microbiology Letters*, **57:** 173-178.

Vierula, P.J. and M. Kapoor (1989). NAD-specific glutamate dehydrogenase of *Neurospora crassa*. *J. Biol. Chem.*, **264:** 1108-1114.

Weinbach, E.C., C.E. Claggett, D.B. Keister, and L.S. Diamond (1980). Respiratory metabolism of *Giardia lamblia*. *J. Parasitol.*, **66:** 347-350.

Weintraub, H., S.J. Tapscott, R.L. Davis, M.J. Thayer, M.A. Adam, A.B. Lassar, and A.D. Miller (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc. Natl. Acad. Sci. U.S.A.*, **86:** 5434-5438.

Yee, J. and P.P. Dennis (1992). Isolation and characterization of a NADP-dependent glutamate dehydrogenase gene from the primitive eucaryote *Giardia lamblia*. *J. Biol. Chem.*, **267:** 7539-7544.

Zomerdijk, J.C., R. Kieft, and P. Borst (1991). Efficient production of functional mRNA mediated by RNA polymerase I in *Trypanosoma brucei*. *Nature*, **353:** 772-775.