Molecular analysis of the congopain gene family

by

Erastus Mulinge Kakundi

B.S.c (Hons) Kenyatta University

Submitted in fulfilment of the

Academic requirements for the degree of

Master of Science

in

Biochemistry

in

School of Biochemistry, Genetics, Microbiology and Plant Pathology

University of KwaZulu-Natal

Pietermaritzburg

December 2008

PREFACE

This thesis is my original work and has not been presented for a degree in any other university or for any other award

Erastus Mulinge Kakundi

Signature _____   Date _____

Supervisor

Prof. Theresa HT Coetzer
Chair Department of Biochemistry
School of Biochemistry, Genetics and Microbiology
University of KwaZulu-Natal Pietermaritzburg

Signature _____   Date _____

Co-supervisor

Dr. Alain Boulangé
Department of Biochemistry
Visiting Scientist from CIRAD, France
School of Biochemistry, Genetics and Microbiology
University of KwaZulu-Natal Pietermaritzburg

Signature _____   Date _____

DEDICATION

I dedicate this work to my beloved wife Irene and our son Befon, whose love and motivation led to the success of this work. To my parents Julius Kiilu and Margaret Kakundi for their concerted efforts to educate me and lastly to my brothers, sisters, relatives and friends whose encouragement and support was enormous.

ABSTRACT

Animal trypanosomosis is a major constraint in livestock production in Sub-Saharan Africa. With the emergence of resistance against trypanocidal drugs, the cost and environmental concerns raised by vector control, and the challenge of antigenic variation in vaccine development, alternative control measures are being sought. An anti-disease strategy, whereby the immune response or chemotherapy is aimed towards pathogenic factors rather than the parasite itself, constitutes such a novel approach. Congopain is the major cysteine protease in *Trypanosoma congolense*, and upon release in the bloodstream of infected cattle, acts as a pathogenic factor. It is therefore an attractive candidate for an anti-disease vaccine. It was hence deemed necessary to investigate the variability of congopain-like cysteine proteases before attempting to design drugs and vaccines based on the inhibition of congopain.

Most congopain-like cysteine protease genes of *T. congolense* exist in a single locus of 12-14 copies organised as tandem repeats of 2 kb gene units. A gene unit library of 120 clones was constructed out of several cosmid clones selected in a previous study that contained various lengths of the congopain locus. Some 24 gene unit clones were sequenced, and it was found that congopain genes cluster in three sub-families, named CP1 (8 clones), CP2 (12 clones) and CP3 (4 clones). The latter most characteristically shows a substitution of the active site cysteine by a serine. Isoform specific primers were designed and used to verify the proportions of the three isoforms (one third CP1, half CP2 and a sixth CP3) in the remaining clones of the library. Since this first study was conducted in one isolate, IL 3000, the results were subsequently validated in a large array of isolates, of *T. congolense*, as well as *T. vivax* and *T. brucei* subspecies, by a PCR approach. Finally, to gain access to copies of congopain genes that are not present in the locus, but rather scattered in the genome, an attempt was made to construct a 2 kb size-restricted genomic library. Only 206 clones could be produced, of which a mere 8 coded for congopain-like proteases. The fact that 7 out of 8 of these clones belong to CP3 (thought to be inactive) suggested a cloning artefact, possibly related to the activity of the cloned proteases.

Overall, all congopain genes appear very conserved in a given species, with 87-99% identity at protein level. The pre- and pro-region were the most conserved, while the catalytic domain was the most variable, especially around the active site cysteine, with frequent replacement by a serine residue, and in one instance by phenylalanine. The histidine residue of the

catalytic triad was also substituted by either a serine or a tyrosine in some instances. The proenzyme cleavage site sequence was also variable, with APEA being the predominant N-terminal sequence. RT-PCR analyses indicated that CP1, CP2 and CP3 mRNA are all present in the bloodstream forms of *T. congolense*, showing that these variants are likely to be expressed.

The conclusion of this study is that, given the high overall conservation of congopain genes in the genome, for the purpose of anti-disease vaccine, it is likely that a single immunogen will suffice to raise antibody able to inhibit all circulating congopain-like cysteine proteases. For chemotherapy however, a more in-depth enzymatic characterisation of the mutants, involving functional recombinant expression, will have to be undertaken.

ACKNOWLEDGEMENTS

CONTENTS

x

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AMV	avian myloblastosis virus
AMV RT	avian myoblastosis virus reverse transcriptase
ATP	adenosine tri-phosphate
*Bam*HI	*Bacillus amyloliquefaciens* H
BLAST	blast local alignment search tool
bp	base pairs
*Bst*ZI	*Bacillus stearothermophilus* Z130
CATT	card agglutination trypanosomosis test
cDNA	complementary deoxyribonucleic acid
CIRDES	Centre International de Recherche-Développement Sur l'Elevage en Zone Subhumide (Burkina-Faso)
cfu	colony forming unit
CRTA	Centre de Recherches sur les Trypanosomoses Animales
DEPC	diethylpyrocarbonate
DNA	deoxyribonucleic acid
dNTP	deoxy nucleotide tri-phosphate
DMSO	dimethylsulfoxide
DDT	dichloro-diphenyl-trichloroethane
dT	deoxythimidine
*Eco*RI	*Escherichia coli* RY 0101
EDTA	ethylenediaminetetraacetic acid
ELISA	enzyme linked immunosorbent assay
EMBL	European Molecular Biology Laboratory
g	relative centrifugal force
GADPH	glyceraldehyde phosphate dehydrogenase
GARP	glutamate-alanine rich proteins
*Hin*dIII	*Haemophilus influenzae* Rd
IgG	immunoglobulin G
IgM	immunoglobulin M
ILRI	International Livestock Research Institute
IPTG	isopropyl-β-D-thiogalactopyranoside

kb	kilobase
kbp	kilobase pairs
KEMRI	Kenya Medical Research Institute
kV	kilo-volt
MCS	multiple cloning site
*Mlu*I	*Micrococcus luteus* I
M-MLV or MuLV	Moloney murine leukaemia virus
mRNA	messenger RNA
*Not*I	*Nocardia otitidis-caviarum*
OD	optical density
ORF	open reading frame
PCR	polymerase chain reaction
PEG	polyethylene glycol
PFGE	pulsed field gel electrophoresis
*Pst*I	*Providencia stuartii* I
RNA	ribonucleic acid
rRNA	ribosomal RNA
RT	room temperature
RT-PCR	reverse transcriptase polymerase chain reaction
*Sal*I	*Streptomyces albus* G
SAP	shrimp alkaline phosphatase
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SERA	serine repeat antigen
SSU rRNA	small-subunit rRNA
TAE	tris acetate buffer
TE	tris EDTA
Tm	melting temperature
VSG	variant surface glycoprotein
*Xba*I	*Xanthomonas badrii* I
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

CHAPTER 1

LITERATURE REVIEW

1.1   Introduction

African trypanosomosis is a disease in man, domestic and wild animals caused by protozoan parasites called trypanosomes. African trypanosomes are generally transmitted to the mammalian hosts by the bite of infected tsetse flies. These parasites live in the blood, lymph nodes, spleen and cerebrospinal fluid of their hosts (Schmidt and Roberts, 1989). Trypanosomes cause sleeping sickness in humans and Nagana in livestock. The symptoms of the disease are chronic anaemia, emaciation, fever, infertility, immunosuppression and finally death. *Trypanosoma brucei rhodesiense* and *T. b. gambiense* are the aetiological agents of Human African Trypanosomosis. *T. b. gambiense* causes chronic infections in humans and is mainly transmitted by riverine tsetse flies belonging to the *Glossina palpalis* species, which is widely spread in West and Central Africa. *T. b. rhodesiense*, which is confined to Eastern and Southern African regions, causes an acute form of the disease.

African animal trypanosomosis (Nagana) is caused by *T. congolense, T. vivax* and to a lesser extent by *T. b. brucei*. The three species are transmitted by tsetse flies, but *T. vivax* can also be transmitted by biting flies (Tabanids and Stomoxes), hence its presence also in South America. *T. congolense* infects cattle, horses, sheep, goats, camels, pigs and dogs. Three groups of *T. congolense* have been characterised, namely Savannah, Riverine/Forest and Kenyan coast/Kilifi (Gashumba *et al*., 1988). *T. evansi,* a parasite closely related to *T. brucei*, but that has lost the ability to be cyclically transmitted, causes Surra in camels, horses, domestic animals and wildlife. It is widely spread all over the world in the intertropical region. In Central and South America *T. evansi* is transmitted by biting flies, but also vampire bats and affects mainly horses. In Asia it affects camels, water buffaloes, horses and pigs, and is mainly transmitted by biting flies. Dourine is caused by *T. equiperdum*, another close relative to *T. brucei*, and is sexually transmitted in equines and camelids. Suids are infected by *T. simiae* and *T. godfreyi*, transmissible by tsetse flies.

It is estimated that 8-11 Million/km^2 of land in sub-Saharan Africa is infested by tsetse flies, a region inhabited by 260-300 million people and 45-50 million cattle (Figure 1.1).

Trypanosomosis impacts negatively on livestock productivity, and this translates into 1.3 to 5 billion US$ worth of losses (McDermott and Coleman, 2001).
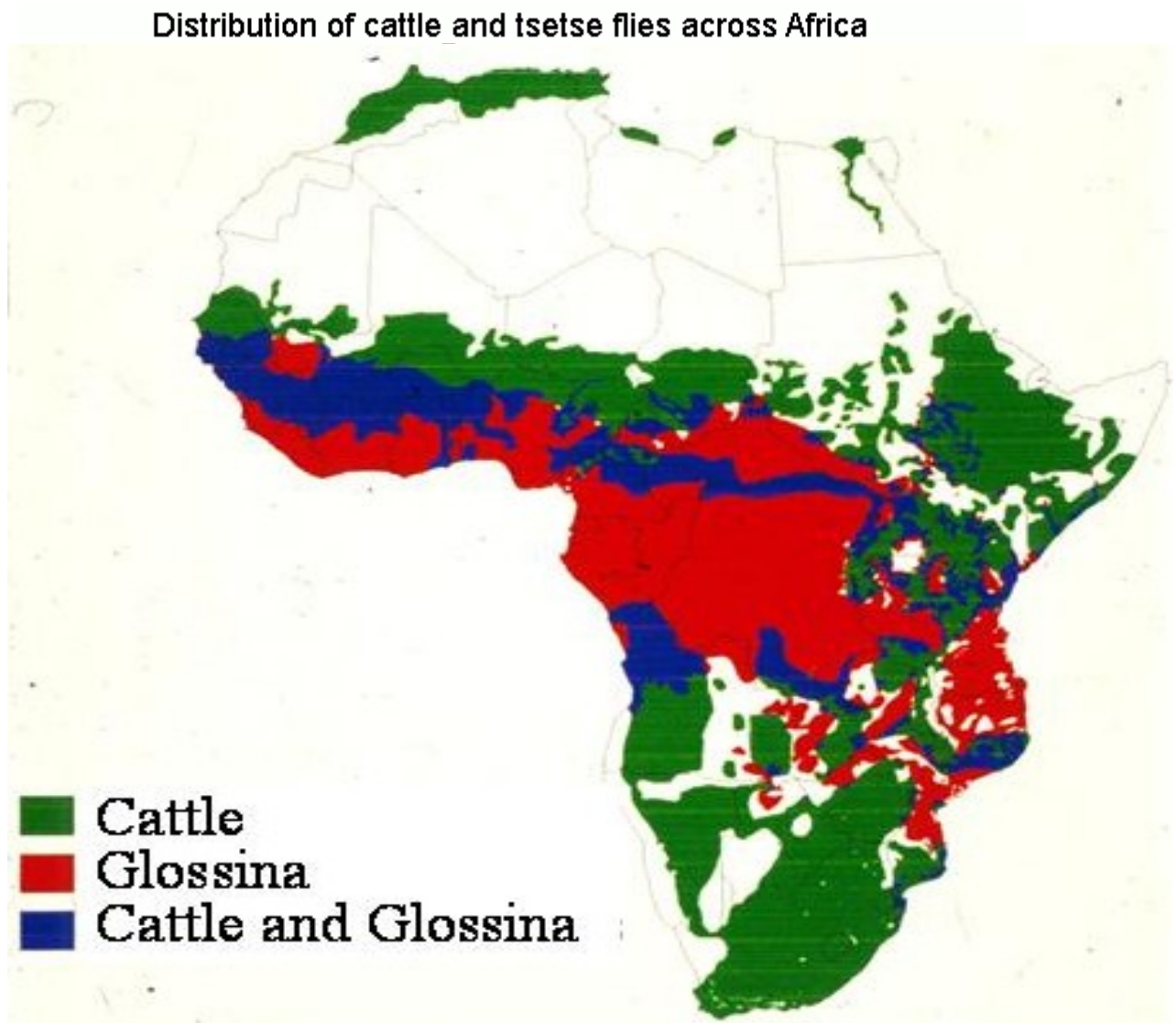
Figure 1.1 Geographical distributions of cattle and tsetse flies across Africa, acquired from University of Liverpool www.genomics.liv.ac.uk/tryps accessed on 01.12. 06

1.2   Morphology of the African trypanosome

The generalised anatomy of African trypanosomes is given in Figure 1.2. Trypanosomes are unicellular organisms which possess a single flagellum arising from the basal body (kinetosome) at the floor of the flagellar pocket which occurs at either the posterior or anterior of the body (Noble and Noble, 1982). The kinetoplast is a sausage or disc-shaped structure containing mitochondrial DNA; it is closely associated with and is posterior to the basal body. The kinetoplast is also important in cyclical transmission by the tsetse fly; it is absent in *T. evansi*, which is not cyclically transmitted. Pellicular microtubules are located

beneath the cell membrane and give structural support to the parasite's body (Schmidt and Roberts, 1989).
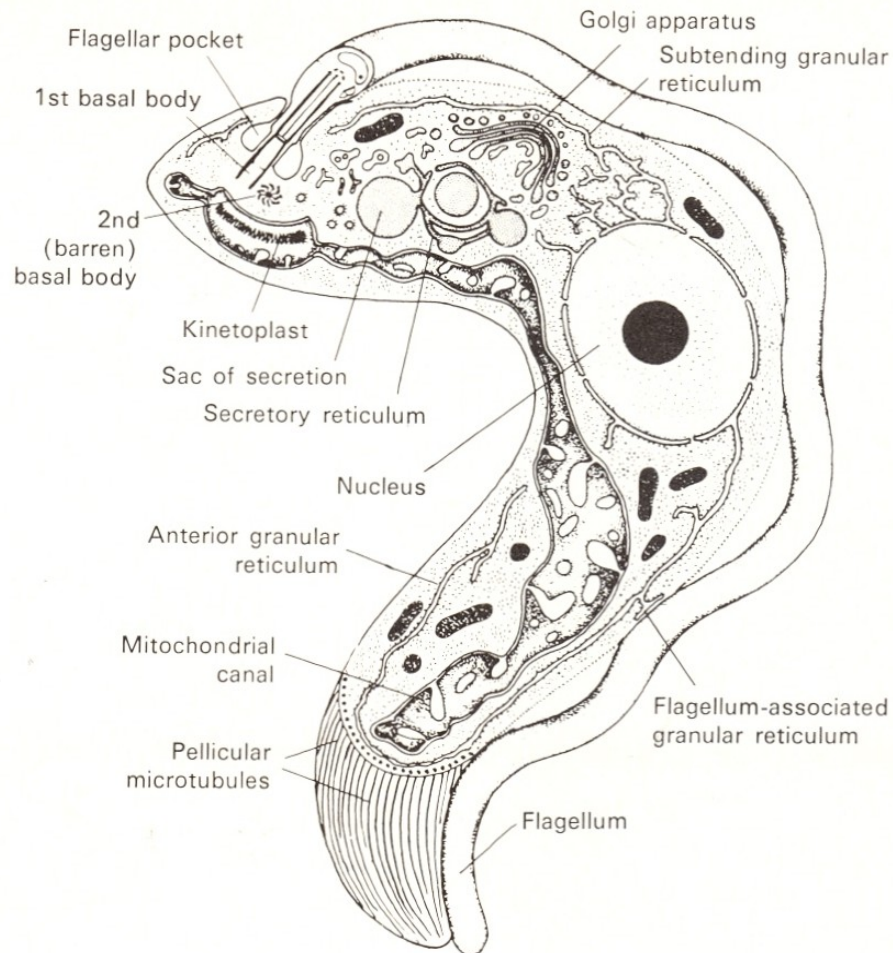
Figure 1.2 Morphology of *Trypanosoma congolense*. The diagram shows structures of the bloodstream trypomastigote form of *T. congolense* as revealed by electron microscopy. [Vickerman, (1969) as acquired from Schmidt and Roberts, (1989)]

1.3   Classification of African trypanosomes

The classification of the genus *Trypanosoma* is given in Figure 1.3. Trypanosomes are unicellular protozoans belonging to the subphylum *Sarcomastigophora* because they possess a single nucleus and use flagella for locomotion. The *Zoomastigophora* class represents organisms without a chloroplast but possessing a flagellum. Members of the order *Kinetoplastida* possess a kinetoplast; in addition their flagellum is attached to the body by an undulating membrane, which is a characteristic of the suborder *Trypanosomatida*. The *Trypanosoma* genus is divided into two groups based on the site of development of the parasite within the insect vector. Salivarian trypanosomes develop in the anterior part of the

alimentary canal of the insect vector, while the stercoraria (like the Chagas' disease agent *T. cruzi*) develop in the posterior part and are present in faeces (Schmidt and Roberts, 1989).

The Trypanosoma genus is believed to be monophyletic (a group of organisms that descend from a single ancestor) (Lukeš *et al*., 1997; Haag *et al*., 1998; Stevens and Gibson, 1999) although previously it had been suggested that this genus is paraphyletic (a taxonomic group which does not include all the descendants of an ancestral taxon) (Gómez *et al*., 1991; Maslov *et al*., 1996). All these observations were based on molecular phylogenetics of 18S Small-Subunit (SSU) rRNA genes. Phylogenetic studies done using genes coding for proteins such as glyceraldehyde phosphate dehydrogenase (GADPH) indicate that *Trypanosoma* is monophyletic (Hannaert *et al*., 1992, 1998; Alvarez *et al*., 1996). *T. vivax* has been shown by phylogenetic analyses to be the first species to diverge from the monophyletic clade of salivarian trypanosomes (Lukeš *et al*., 1997; Stevens and Gibson, 1999; Simpson *et al*., 2006). The subgenus *Nannomonas* consists of *T. congolense, T. simiae and T. godfreyi* and is considered monotypic (a taxonomic group with only one subgroup at the next taxonomic level) based on phylogenetic analyses of 18S SSU and 28 Large-Subunit (LSU) rRNA sequences (Lukeš *et al*., 1997; Hagg *et al*., 1998; Stevens and Gibson, 1999). *T. congolense* is also monotypic (Hagg *et al*., 1998) and its three subgroups (Savannah, Forest and Kilifi) are genetically distinct (Stevens and Gibson, 1999), however, a study on glutamate-alanine rich proteins (GARP) showed conservation of this gene among Savannah, Forest and Kilifi subgroups of *T. congolense* (Asbeck *et al*., 2000). *T. simiae* and *T. godfreyi* show a distinct evolutionary divergent lineage (Gibson *et al*., 2001). GAPDH phylogenetic studies indicated that *T. b. brucei* is monotypic. Moreover based on 18S rRNA *T. b. brucei* was shown to be closely related to *T. b. rhodesiense*, whereas *T. evansi, T. equiperdum* and *T. b. gambiense* cluster together (Stevens and Gibson, 1999).

1.4 Life cycle of African trypanosomes

The life cycle of *T. brucei* is illustrated in Figure 1.4. The short stumpy non-dividing trypomastigotes are acquired in the tsetse fly blood meal. This form of the parasite moves to the midgut and later to the hindgut as a procyclic form. The procyclic forms undergo developmental stages and migrate to the salivary glands, where they transform into epimastigotes, which attach themselves to the salivary gland epithelial lining. Epimastigotes transform into non-dividing metacyclic forms that are infective and located in the lumen of the salivary glands. The metacyclic forms are passed to a new host during a blood meal

(Schmidt and Roberts, 1989). The life cycle of *T. congolense* and *T. vivax* is somewhat simpler, as no stumpy forms are visible, and the maturation stages (epimastigote and metacyclic) occur essentially in the labrum.

1.5   Pathology of African animal trypanosomosis

Metacyclic forms of trypanosomes enter the skin of the host when an infected tsetse fly feeds on the host (blood meal). Here they multiply, causing skin inflammation and lesions called chancre, and changes in the mast cells. The parasites proceed to the lymph nodes, characterised by increased levels of CD4 and CD8 cells, higher ratios of lymphoblast and B cells and enlargement of lymph nodes. Two to three weeks after infection, trypanosomes are detected in the blood and by 3-4 weeks antibodies against parasite antigens are present in the circulation. IgG1, IgG2 and IgM for variant surface glycoprotein (VSG) epitopes are detected in *T. congolense* infected cattle, but fail to be protective (Williams *et al.*, 1996). Depletion of C3 of complement and immunosuppression are also observed in trypanosome infected cattle (Authié and Pobel, 1990). There is a marked reduction in leukocytes (leukopaenia) and circulating platelets as well as anaemia. Erythrocyte destruction in the spleen, liver, lymph nodes, lungs and bone marrow contribute to anaemia in both the acute and chronic phases. Other clinical features of trypanosomosis include heart damage, infertility, low milk production, reduced capacity to do work, abortion, poor growth and weight loss (Murray *et al.*, 1991).

1.6   Genome of African trypanosomes

Trypanosomes have two types of genomes, namely the nuclear and kinetoplast genomes. The nuclear genome is diploid; however some parts of the genome are haploid. In *T. b. brucei* the haploid nuclear genome is about 35 Mb and it is thought to contain as many as 12000 genes (El-Sayed *et al.*, 2000). Since the nuclear chromosomes of *T. b. brucei* have been shown not to condense during mitosis, they have been characterised by pulsed field gel electrophoresis (PFGE). When separated by PFGE, three classes of chromosomes can be observed, megabase chromosomes, intermediate chromosomes and minichromosomes. Megabase chromosomes are also called house-keeping chromosomes and account for approximately 80% of the nuclear genome. They contain most of the genes necessary for cellular activities of trypanosomes. There are 11 megabase chromosomes ranging from 1 to 6 Mb in size and are diploid in the nucleus (Melville *et al.*, 1998, 2000). The megabase chromosomes are assigned numbers I-XI according to their sizes (Turner *et al.*, 1997; El-Sayed *et al.*, 2000) from the smallest to the largest. There is a significant degree of size

Subkingdom   Protozoa

Phylum   Sarcomastigophora

Subphylum   Mastigophora

Class   Zoomastigophora

Order   Kinetoplastida

Family   Trypanosomatida

Genus   *Trypanosoma*

Section   Stercoraria   Salivaria

Subgenus   Schizotrypanum   Megatrypanum   Herpetosoma   Duttonella   Nannomonas   Trypanozoon   Pycnomonas

Species   *T. rangeli*   *T. cruzi*   *T. theileri*   *T. lewisi*   *T. vivax*   *T. congolense*   *T. simiae*   *T. godfreyi*   *T. brucei*   *T. evansi*   *T. equiperdum*   *T. suis*

Subspecies *T. r. Tereja*   *T. c. marinkellei*   *T. v. ziemann*   *T. v. viennei*   *T. b. brucei*   *T. b. rhodesiense*   *T. b. gambiense*

Figure 1.3 Classification of trypanosomes. The illustration shows the classification of trypanosomes based on 18S SSU ribosomal RNA gene sequence analysis (Stevens and Gibson, 1999; Stevens *et al*., 2001).
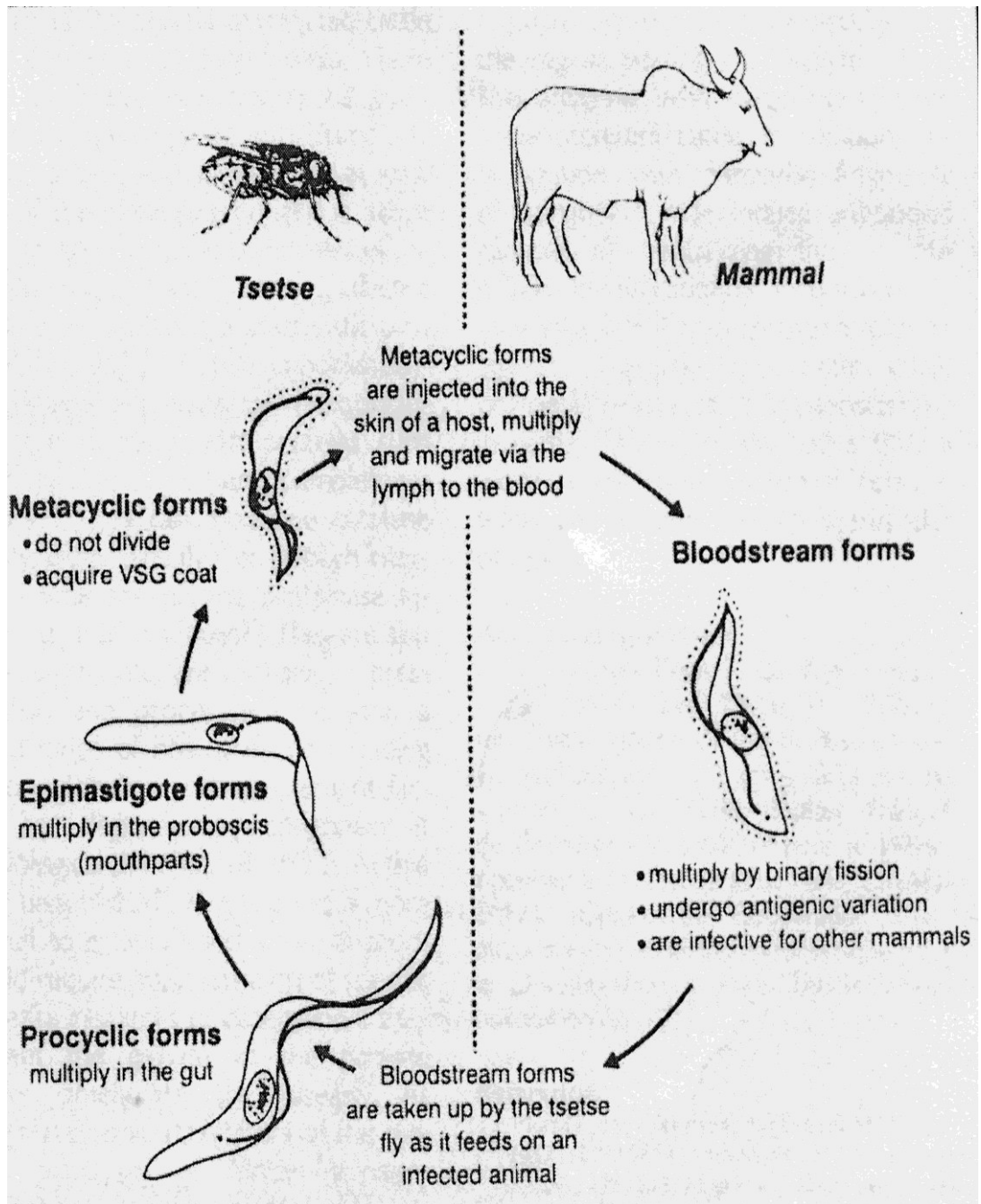
Figure 1.4 Life cycle of African trypanosomes (Authié, 1994)

polymorphism in homologous chromosomes according to strains (Melville *et al*., 1998, 2000). The intermediate chromosomes are 100-700 kb and do not posses the 177 bp repeated sequence of minichromosomes. There are about one to five intermediate chromosomes accounting for 0.2-1% of the nuclear genome. They contain VSG expression sites, telomeric sequences and some gene families. The mini-chromosomes are monoploid, 50-100 kb in size and account for 20% of nuclear DNA content. There are approximately 100 linear

minichromosomes, consisting of a 177 bp repetitive sequence, telomere repeats and a few variant surface glycoprotein genes (Weiden *et al*., 1991).

Two classes of circular DNA structures characterise the kinetoplast genome, the maxicircles and minicircles. Maxicircles are 25-27 kb in size and a single cell contains 50 maxicircles. Maxicircles are composed of 9S and 12S ribosomal RNA, 17 open reading frames (ORF) and 8 kb regions of repetitive sequences. The ORF encodes for several genes of known function, clustered and without introns. The minicircles are approximately 1 kb in size and 5000-10000 minicircles make up a cell. Minicircles are composed of 3-4 guide RNAs (for editing) (Stuart *et al*., 1997), 18 bp inverted repetitive sequences, and one origin of replication. Unlike maxicircles, the minicircles do not encode any protein (El-Sayed *et al*., 2000).

1.7   Diagnosis of African trypanosomosis

For successful treatment and control of African trypanosomosis, cost effective, quick and reliable diagnostic techniques are required. Numerous methods have been used for the diagnosis of African animal trypanosomosis and include observation of clinical signs, parasitological, immunological and molecular based techniques.

1.7.1   Clinical diagnosis

Clinical diagnosis is the physical examination of the clinical signs of the disease. However, this method is not effective as the clinical signs are not specific for trypanosomosis (Rebeski *et al*., 1999). After diagnosis, infected animals are treated with trypanocides and recovery is a true confirmation of the diagnosis.

1.7.2   Parasitological methods

Detection of trypanosomes in wet blood films and Giemsa-stained blood films requires a high level of parasites in blood; this method has been improved by concentrating the parasites by centrifugation techniques i.e. the haematocrit centrifugation technique and the buffy coat technique. These two concentrating techniques were found to be successful in the detection of trypanosomes and, with highly trained personnel, to give similar sensitivity as PCR (Desquesnes, 1997).

1.7.3   Immunological methods

Several immunological methods have been developed for the diagnosis of trypanosomosis. The complement fixation test has been found successful in the diagnosis of *T. equiperdum* especially in countries where other members of the subgenus *Trypanozoon* are not found

(Williamson *et al*., 1988). The indirect fluorescent antibody test (IFAT) is limited in that it is not species-specific and cannot be used to determine the severity of the infection. Another reliable method for diagnosis is the card agglutination trypanosomosis test (CATT); it can be used in the field where most methods cannot be applied. CATT has been used in the diagnosis of *T. b. gambiense* sleeping sickness and in *T. evansi* infections in animals. Enzyme linked immunosorbent assay (ELISA) is a very reliable method for diagnosis because it does not require highly specialised equipment and can be used in the field. Although ELISA is a common diagnostic method, it is limited to the detection of anti-trypanosomal antibodies and cannot determine a current or past infection as antibodies persist for a longer time than parasites within the host (Van den Bossche *et al*., 2000).

1.7.4   Molecular based methods

Molecular methods employed in the diagnosis of African trypanosomosis are based on the detection and/or amplification of nucleic acids. One of these methods exploits the existence of a 177 bp repetitive sequence in the trypanosome genome; a set of six primers has been used for the differentiation of members of the *Nannomonas* subgenus (Masiga *et al*., 1992, 1996). Kinetoplast DNA (kDNA) has also been used as a probe in identifying trypanosomes because their DNA is conserved. This probe is however limited in that some trypanosome species do not have kDNA. A few species-specific diagnostic tools have been developed; one of these is based on GARP and is characteristic of the *Nannomonas* subgenus (Asbeck *et al*., 2000). Welburn *et al*., (2001) developed a diagnostic technique based on the serum-resistance associated (SRA) gene to differentiate between *T. b. brucei* and *T. b. rhodesiense*. Other diagnostic methods that have been developed include restriction fragment length polymorphism (RFLP; Hide *et al*., 1994), randomly amplified polymorphism DNA (RAPD) and amplified fragment length polymorphism (AFLP) (Masiga *et al*., 2000). Lately, ITS (Internal Transcribed Spacer of ribosomal DNA repeating units) have been used for the species-specific diagnostics of trypanosome. It allows with one set of primers to distinguish most of African trypanosome species in a single PCR reaction, based on the size-polymorphism (Desquesnes et al. 2001). These methods require large quantities of DNA, highly qualify personnel, and expensive equipment, and hence are usually used for epidemiological studies rather than diagnosis.

1.8   Control of African trypanosomosis

Development of a vaccine has been hampered by the ability of trypanosomes to alter their VSG through the mechanism of antigenic variation (Donelson, 2003). Use of trypanocides is

expensive, the drugs are toxic and trypanosomes have become resistant to trypanocides. Vector control methods involve bush clearing, elimination of wildlife, and the use of insecticides, traps and screens. Successful control of trypanosomosis has to involve both parasite and vector control, in an integrated manner.

1.8.1 Bush clearing and elimination of wildlife

Bush clearing involves destruction of tsetse flies habitat and serves as a physical barrier to prevent reinvasion. The method is expensive, difficult to maintain and ecologically unsound. An improved method of bush clearing is agronomic prophylaxis, whereby bush clearing is followed by planting of crops (Pépin and Méda, 2001). Elimination of wildlife, that removes both tsetse attractant and parasite reservoir, proved to be a very effective method of control, but an ecological disaster. It was terminated based on ethical and environmental considerations (Schofield and Maudlin, 2001).

1.8.2 Use of insecticides

Dichloro diphenyl trichloroethane (DDT), dieldrin, endosulfan and synthetic pyrethroids are insecticides that have been used for the control of trypanosomosis. DDT was in use some decades ago and was very effective for tsetse control. However due to its perceived high level of toxicity its use was prohibited by many countries. Less harmful synthetic pyrethroids replaced the toxic organochlorides, as they are metabolised by mammals. In Nigeria insecticides were successfully used to eliminate tsetse fly in an area of approximately 200,000 km^2 (Schofield and Maudlin, 2001). Impregnated traps and screens attract tsetse flies because of their shape, colour or contrast between colours.

1.8.3 Trypanotolerance

Cattle are said to be trypanotolerant because they are able to control the pathogenic effects of the disease and control parasite proliferation (Murray and Dexter, 1988). Trypanotolerance is a genetic trait of *Bos taurus* breeds of West and Central Africa, particularly the N'Dama longhorn and the West African short horn breeds. Cattle without this trait are called trypanosusceptible, which include most *Bos indicus* (Zebu) breeds. Trypanotolerant cattle are able to control anaemia and parasitaemia, hence remain productive under tsetse pressure (Murray *et al*., 1991; d'Ieteren *et al*., 1998,). Cattle breeds with this trait are documented to also possess the ability to resist other infectious diseases such as tick-borne illness (Murray *et al*., 1982; Mattioli *et al*., 2000). They are however of small size and have never been adopted out of their region of origin. Some breeding program have been attempted to transfer trypanotolerance to more conventional breeds.

1.8.4 Use of trypanocides

Over the last 50 years three trypanocides have been used for treating bovine trypanosomosis in Africa. They include isometamidium chloride, homidium (chloride and bromide) and diminazene aceturate. Isometamidium compounds have prophylactic properties while homidium have both prophylactic and therapeutic effects. Diminazene is used only as a therapeutic agent. No new compounds have been developed due to the high cost of developing and licensing new compounds and also the limited market in Africa. As a result many African pharmaceutical companies have developed cheap generic forms of the trypanocides, which are assumed to be of low quality. Due to prolonged use of trypanocides, resistance to one or more of these compounds has been observed in African trypanosome-affected countries (Geerts *et al*., 2001). A recent study on the efficacy of antibiotic ascofuranone on *T. vivax* infected mice showed that the antibiotic has chemotherapeutic properties (Yabu *et al*., 2006). There is an urgent need to develop new chemotherapeutic targets. Cysteine protease inhibitors have been shown to possess chemotherapeutic properties against trypanosomosis (McKerrow, 1999; McKerrow *et al*., 1999, Caffrey *et al*., 2000, Lalmanach *et al*., 2002). The main problem resides in designing parasite-specific inhibitors, as proteases are often phylogenetically conserved.

The same cysteine proteases are also targets for vaccine design. A hybrid protein vaccine consisting of leishmanial major cysteine proteinases Type I (CPB) and Type II (CPA) was found to give partial protection against leishmaniasis (Zadeh-Vakili, *et al*., 2004). In trypanosomosis, congopain, the major cysteine protease, has been shown as a potential vaccine candidate (Authié *et al*., 2001, Lalmanach *et al*., 2002). The knowledge on the role and relatedness of congopain and other related cysteine proteases is paramount in the development of both vaccines and chemotherapeutic drugs against trypanosomosis.

1.9 Classification of proteases

Proteases catalyse the breakdown of peptides and proteins by hydrolysing peptide bonds. They are classified into two major classes, i.e. endopeptidases and exopeptidases (Barrett, 1994). Exopeptidases catalyse the sequential removal of amino acid residues from the amino-(aminopeptidases) or carboxy-(carboxypeptidases) termini of peptides or proteins. Endopeptidases catalyse cleavage of peptide bonds within a protein. Proteinases are further classified into five distinct groups depending on the amino acid residues that make up the

catalytic unit, namely cysteine, serine, aspartic, metallo- and threonine peptidases (Barrett, 1994; Travis and Potempa, 2000; Rawlings and Barrett, 1999, 2000).

The essential amino acid residues in the active site of cysteine proteases are cysteine, histidine and asparagine (Storer and Ménard, 1994). Serine, aspartate and histidine form a catalytic triad in serine proteases (Rawlings and Barrett, 1994). Catalysis by serine proteases occur after formation of a tetrahedral transition state intermediate. Eukaryotes have a number of these serine proteases such as trypsin, chymotrypsin and oligopeptidase B (Coetzer *et al*., 2008). Two aspartate residues form the catalytic triad in aspartic proteases (Rawlings and Barrett, 1995b). Common aspartic proteases include the digestive enzymes pepsin and rennin, and HIV-1 protease. Metalloproteases are a group of enzymes which require metal ions such as zinc for enzymatic activity (Rawlings and Barrett, 1995a). Matrix metalloproteases that hydrolyse collagen and gelatine are such enzymes found in eukaryotes. The proteasome (Rivett, 1993) is a threonine protease where the primary catalytic group is the hydroxyl of threonine. It is a cytoplasmic complex of multiple subunits which have three independent hydrolytic activities, i.e. trypsin-like, chymotrypsin-like and peptidylglutamyl-peptide hydrolase activities (Rivett, 1993, Seemüller *et al*., 1995).

1.10   Proteases of protozoan parasites

Proteases have diverse roles in protozoan parasites, namely invasion of host cells and tissues, hydrolysis of host proteins, stimulation of host immune responses and involvement in differentiation (Rosenthal, 1999; Klemba and Goldberg, 2002; McKerrow *et al*., 2006). In African trypanosomes serine oligopeptidase B from *T. b. brucei* plays a major role in trypanosomosis pathogenesis through the degradation of peptide hormones in the bloodstream of infected hosts (Troeberg *et al*., 1996; Morty *et al*., 1999; Coetzer *et al*., 2008). Oligopeptidase B from *T. evansi* inactivates atrial natriuretic factor in the bloodstream of infected hosts (Morty *et al*., 2005). Recently an aminopeptidase pyroglutamyl peptidase type I, from *T. b. brucei* has been shown to regulate the levels of peptide hormones such as gonadotropin-releasing hormone and thyrotrophin-releasing hormone in the plasma of infected host (Morty *et al*., 2006).

Serine proteases of merozoites of *Plasmodium falciparum* are involved in host cell invasion (Blackman and Holder, 1992). Serine-rich antigen (SERA) is a group of cysteine proteases of *Plasmodium* species which possess a central domain that show homology to the papain

family of cysteine proteases, but exhibit an unusual cysteine to serine substitution at the active site of cysteine residue (Bzik *et al*., 1988; Knapp *et al*., 1991; Kiefer *et al*., 1996; Gor *et al*., 1998; Hodder *et al*., 2003).

In *Leishmania amazonensis* zinc metalloprotease leishmanolysin (GP63) is a major surface glycoprotein in the infective metacyclic promastigote that is involved in invasion of macrophages (Frommel *et al*., 1990; McGwire and Chang, 1994,). This role was investigated in *L. amazonensis* by the expression of a *gp63* antisense construct and led to impaired parasitisation of macrophages (Chen *et al*., 2000). GP63 has also been associated with resistance against host complement degradation as shown in *L. major*, where gp63 mutant parasites were more susceptible to lysis by human serum as well as delayed pathogenesis in susceptible mice (Joshi *et al*., 2002).

Arginine aminopeptidases are necessary for excystation of *Cryptosporidium parvum* (Okhuysen *et al*., 1994). Two aspartic proteases are associated with haemoglobin catabolism in *P. falciparum* (Goldberg *et al*., 1991), a metalloprotease called falcilysin has the same role in *P. falciparum* (Eggleson *et al*., 1999).

1.11 Cysteine proteinase

Cysteine proteases belonging to the papain family have been found in a wide range of organisms including bacteria, plants, invertebrates and vertebrates (Berti and Storer, 1995). They are the major proteolytic enzymes in protozoan parasites and important virulence factors (Sajid and McKerrow, 2002). In *Leishmania*, cysteine proteases are responsible for the destruction of histocompatibility complex class II molecules in *L. amazonesis* (De Souza Leao *et al*., 1995). Cysteine proteases have been characterised in *L. pifanoi* (Traub-Cseko *et al.,* 1993), *L. major* (Sakanari *et al.,* 1997), *L. donovani chagasi* (Omara-Opyene and Gedamu, 1997), *L. mexicana* (Mottram *et al*., 1992), *L. donovani* complex (Mundodi *et al*., 2002) and *L. ethiopica* (Kuru *et al*., 2007).

Trypanopain is the major cysteine protease in *T. b. brucei* (North *et al*., 1983; Lonsdale-Eccles and Mpimbaza., 1986; Lonsdale-Eccles and Grab, 1987; Robertson and Coombs, 1990; Troeberg *et al*., 1996). Trypanopain is a lysosomal cysteine protease of the papain-family (Lonsdale-Eccles and Grab, 1987) that plays a major role in the differentiation of bloodstream-form trypanosomes (Pamer *et al*., 1989). Cruzipain (Cazzulo *et al*., 1989, 1990),

the major protease in *T. cruzi* is expressed in all life cycle stages (Eakin *et al*., 1992; Tomás and Kelly, 1996).

A family of four papain-family cysteine proteases called falcipains has been characterised in *P. falciparum* (Sijwali *et al*., 2006), they include falcipain-1 (Rosenthal and Nelson, 1992), falcipain-2 (Shenai *et al*., 2000), falcipain-2' (Singh *et al*., 2006) and falcipain-3 (Sijwali *et al*., 2001). Falcipain-2 and falcipain-3 are associated with haemoglobin degradation in trophozoites (Shenai *et al*., 2000; Sijwali *et al*., 2001; Dahl and Rosenthal, 2005). Falcipain-1 has been associated with erythrocyte invasion in *Plasmodium* species (Mayer *et al*., 1991). The biological role of falcipain-2' still remains unknown (Singh *et al*., 2006).

In *Entamoeba histolytica* cysteine proteases are potential virulence factors (Olivos-Garcia *et al*., 2004; Bruchhaus and Tannich, 1996) and play a role in adhesion of trophozoites to the mucosal lining (Garcia-Rivera *et al*., 1999). Bruchhaus *et al*., (2003) reported a total of 20 cysteine protease genes in *E. histolytica* with only EhCP1, EhCP2 and EhCP5 possessing proteolytic activity.

Three cysteine proteinases of the papain family have been described in *T. congolense*, they are CP1 (Fish *et al*., 1995, EMBL accession number Z25813), CP2 (Jaye *et al*., 1994, EMBL accession number L25130) and CP3 (Downey and Donelson, 1999 EMBL accession no. AF139913). The proteases predicted by the gene sequence of CP1 and CP2 are identical in their pre and pro-peptides, differing with CP3 by only two amino acid residues at the signal peptide, their C-terminal extension shows 86-95% identity, while their catalytic domains are 88-95% identical. However they differ in their N-terminal sequences APPA for CP1 and CP3, APEA for CP2. CP3 differs from CP1 and CP2 by a replacement of a cysteine residue at position 25 (papain numbering) by a serine residue.

1.12 Congopain

Congopain is a 33 kDa cysteine proteinase purified from *T. congolense* bloodstream form lysates (Authié *et al*., 1992). Congopain is made up of a prepeptide followed by a hydrophilic propeptide, which is cleaved to give the mature enzyme consisting of the catalytic or central domain and the C-terminal extension (Figure 1.5). The catalytic domain is linked to the C-terminal extension through a polyproline hinge. Congopain shares structural and functional resemblance with cruzipain from *T. cruzi* and with mammalian cathepsin L

(Lalmanach *et al.*, 2002). The C-terminal extension is trypanosome-specific and its function remains unknown (Stoka *et al.*, 2000). The C-terminal extension has been shown to be unnecessary for enzyme activity (Eakin *et al.*, 1993; Boulangé *et al.*, 2001), enzyme stability and trafficking to the lysosomes (Huete-Pérez *et al.*, 1999; Stoka *et al.*, 1998, 2000). It has been suggested that due to its high antigenicity, it might divert the immune response from the functional region (Lalmanach *et al.*, 2002).

Congopain shows preference for substrates with arginine at P_1, and bulky hydrophobic residues at the P_2 position (Chagas *et al.*, 1997) and a marked preference for a proline residue at P_2' and P_2 (Lalmanach *et al.*, 2002). The fluorogenic peptide Z-Phe-Arg-NHMec, is routinely used to assay for activity. Congopain is inhibited by E-64, peptidyldiazomethanes and cystatin (Mbawa *et al.*, 1992; *Chagas et al.*, 1997), but forms complexes with natural inhibitors such as kininogen and α-macroglobulin while retaining its enzymatic activity (Lalmanach *et al.*, 2002).

Authié (1994) observed that trypanotolerant cattle such as the N'Dama breed (*Bos taurus*) developed a higher IgG response to congopain than trypanosusceptible cattle such as the African Zebu breeds (*Bos indicus*) during primary infection with *T. congolense*, suggesting that congopain may play a role in pathology of trypanosomosis. Furthermore after immunising cattle with CP1/CP2, it was also observed that anti-CP antibodies play a role in modulating trypanosome induced-pathology (Authié *et al.*, 2001). In order to develop an anti-disease vaccine for trypanosomosis based on congopain and other related cysteine proteases it is necessary to study them as described in the present study.
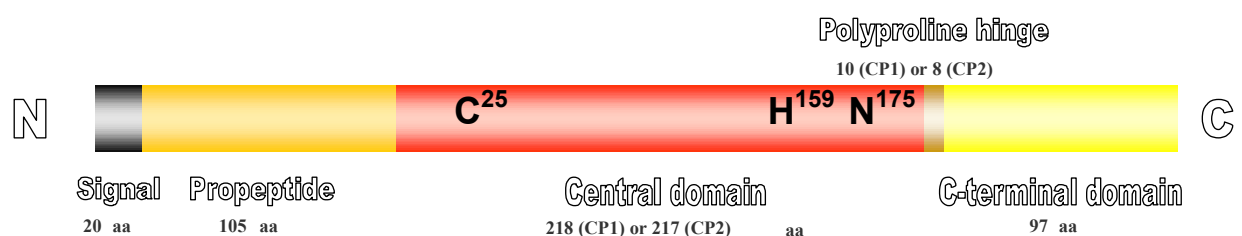
Figure 1.5 Structure of congopain. The figure shows the congopain signal peptide, propeptide, central domain, polyproline hinge and C-terminal domain.

1.12.1 Organisation of congopain-like genes in *T. congolense*

The genomic organisation of congopain was studied by Southern blotting with the congopain coding sequence as a probe (Boulangé *et al.*, manuscript in preparation). *T. congolense* IL 3000 genomic DNA was digested by *Eco*RI and *Hin*dIII restriction enzymes that do not cut within the genes. One strong band at approximately 25 kb was observed, suggesting that the gene exists at a single locus. When the genomic DNA was cut by *Bam*HI, *Pst*I and *Aat*I, which cut once within the gene, a strong band on the blot was observed at approximately 2 kb, suggesting the size of the gene units. A partial digest by enzymes that cut once within the gene demonstrated that 12-14 copies existed in the genome in a tandem repeat (Boulangé *et al.*, in preparation), a similar tandem repeat (19 copies) of cysteine protease genes (*lmcpb*) from *L. mexicana* has been reported (Souza *et al.*, 1992). The *T. b. brucei* genome has 20 copies of cysteine protease genes tandemly arranged (Mottram *et al.*, 1989) and in *T. cruzi* 6 copies of cysteine protease genes are also arranged in a tandem array (Eakin *et al.*, 1992). It was also observed that longer exposure time of the blots resulted in other bands apart from the band designating the congopain gene locus, this led to the hypothesis that other congopain-like genes could be scattered in the genome but outside of the main locus.

A better understanding of other congopain-like cysteine proteases in trypanosomes is required for the design of chemotherapeutic agents and the development of an anti-disease vaccine for trypanosomosis that will inhibit the activities of congopain and other related cysteine proteases.

1.13   Objectives of the study

Congopain has been shown to contribute to the pathology of trypanosomosis (Authié *et al.*, 1993, 2001; Authié, 1994; Boulangé *et al.*, 2001). While an initial immunisation/challenge experiment (Authié et al., 2001) showed a promising use of congopain as an antigen for the design of an anti-disease vaccine, subsequent trials were disappointing (Boulangé, *pers. comm.*). One hypothesis that was put forward is that congopain exists as many variants, and inhibiting only one of these variants, such as CP2, may not be sufficient. Indeed, three papain-like cysteine proteinase genes have been described in *T. congolense* (Fish *et al.*, 1995; Jaye *et al.*,1994; Downey and Donelson, 1999). For the development of an anti-disease vaccine based on congopain that will inhibit all congopain-like cysteine proteases, a study of their variability in trypanosomes is fundamental. Thus the main objective of the present study

was to determine the variability of congopain-like cysteine protease genes in trypanosomes infective to livestock (*T. congolense*, *T. vivax* and *T. b. brucei*).

The first variability study on congopain-like cysteine protease genes was done by studying the clones from a library of 2 kbp congopain gene units, constructed from clones selected in a previous study from a *T. congolense* IL 3000 cosmid library (Boulangé et al., manuscript in preparation). In Chapter 3, the construction of the library by digestion with *Pst*I of 15 positive cosmid clones and cloning of the released 2 kb gene units is described. The variability of the clones was assessed by sequencing, and study of the polymorphism by PCR.

In Chapter 4 a description is given of how the genetic diversity of congopain-like cysteine protease genes across strains and species was studied. Coding and intergenic regions of congopain-like cysteine protease genes from *T. congolense, T. vivax, T. b. brucei*, *T. simiae* and *T. equiperdum* were amplified by PCR using degenerate and non-degenerate primers. The amplified fragments were cloned and sequenced. The DNA sequences were translated into predicted amino acid sequences and analysed.

A genomic organisation study of congopain genes prior to the reported study had revealed the existence of other congopain-like cysteine protease genes outside of the main locus and scattered in the genome (Boulangé et al., manuscript in preparation). Getting access to these copies is the objective covered in Chapter 5. It was reasoned that by constructing a 2 kbp size-restricted subgenomic library, the scattered congopain-like cysteine protease genes would be included in the library. Cloning of the 2 kb fragment obtained from digesting *T. congolense* IL 3000 genomic DNA with *Mlu*I, one of the very rare restriction enzymes that cut within the intergenic region, required the construction of a cloning vector with an *Mlu*I restriction site.

Another general objective of the study was to determine whether the different variants, or family of variants of congopain-like cysteine protease genes are all expressed in trypanosomes. RT-PCR was conducted on purified mRNA with primers targeting different variants of congopain-like cysteine protease genes. These results are also reported in Chapter 5, since the small number of results obtained did not warrant a separate Chapter.

CHAPTER 2

GENERAL MATERIALS AND METHODS

2.1   Introduction

A variety of common molecular biology methods and bioinformatics tools were used throughout the present study and are described in this Chapter. Specifics of every experiment are described in their appropriate sections.

2.2   Suppliers of materials

The suppliers of molecular biology kits (Table 2.1), restriction and other enzymes (Table 2.2) are detailed in the indicated tables, while primers are detailed in the respective Chapters. Bovine serum albumin (BSA), 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal), molecular weight markers, 6× loading dye, analytical grade agarose, ethidium bromide and adenosine tri-phosphate (ATP) were purchased from Promega (USA) and Fermentas life sciences (Commercialised by Inqaba, South Africa). One Shot® Top10 competent cells were from Invitrogen life technologies (USA). Ampicillin and isopropyl-β-D-thiogalactopyranoside (IPTG) were from Roche (Germany). Standard chemicals of the purest grade available were purchased from Merck or Sigma (South Africa). Genomic deoxyribonucleic acid (DNA) samples (Table 2.3) were provided by the International Livestock Research Institute (ILRI), Nairobi, Kenya and the CIRDES.

Table 2.1 Molecular biology kits used in the present study

Name	Use	Supplier
Wizard® *Plus* SV Minipreps DNA purification system	Isolation of plasmid DNA (Section 2.11)	Promega
Wizard® SV Gel and PCR clean-up system	DNA cleaning/purification (Section 2.6)	Promega
Wizard® DNA clean-up system	DNA cleaning/purification (Section 2.12)	Promega
Gene editor *in vitro* site-directed Mutagenesis system	Site-directed mutagenesis (Section 5.2.2.3, 5.2.4, 5.2.5)	Promega
pGEM®-T Easy Vector System I (with competent *E. coli* JM 109 cells)	Cloning of PCR products and transformation (Sections 2.7, 2.9.5)	Promega
TransformAid™	Preparation and use of competent cells and transformation (Sections 2.8.1, 2.9.1)	Fermentas
Genomic DNA purification	Extraction of genomic DNA (Section 5.4.2)	Fermentas
Rapid ligation	Ligation (Section 5.2.9)	Fermentas
AmpliTaq® DNA Polymerase with GeneAmp® PCR kit	Amplification of genomic DNA/colonies (Sections 2.4, 2.10, 3.2.7, 4.3, 4.4.4, 5.6, 5.8)	Applied Biosystem
PCR nucleotide mix	PCR (Sections 2.4, 2.10, 3.2.7, 4.3, 4.4.4, 5.6, 5.8)	Roche
SV Total RNA Isolation	Isolation of total RNA (Sections 2.5, 5.8.2)	Promega
Access RT-PCR	Amplification of mRNA (Section 2.5)	Promega
RNAid®	Isolation and purification of RNA (Sections 2.5, 5.8.2)	Bio 101 systems
One Shot® TOP10 Chemically Competent Cells	Transformation (Section 5.5.2)	Invitrogen

Table 2.2 Enzymes used in the present study, their purpose and supplier

Name	Function	Supplier
Mlu I	Restriction digestion (Sectio 5.5.1)	Promega/Fermentas
*Eco*RI	Restriction digestion (Sectio 5.2.5)	Promega/Fermentas
*Pst*I	Restriction digestion (Section 3.2.1, 3.2.2, 3.2.6, 5.7)	Promega/ Boehringer Mannhem
*Sal*I	Restriction digestion (Section 5.7	Promega/ Boehringer Mannhem
*Xba*I	Restriction digestion (Sectio 5.2.5)	Promega
T4 polynucleotide kinase	Phosphorylation (Section 5.2.8)	Promega
T4 DNA ligase	Ligation (section 5.2.9)	Fermentas
Shrimp alkaline phosphatase	Dephosphorylation (Section 3.2.3, 5.3, 5.6)	Roche
Proteinase K	Hydrolysis of proteins (Sectio 5.4.1)	Sigma
RNases	Hydrolysis of RNA (Section 5.4.1, 5.4.2)	Sigma

Table 2.3 Genomic DNA from different trypanosomal species and strains obtained from CIRDES and ILRI

Lab assigne No.	Name of species/strains	Lab assigne No.	Name of species/strains
Tl1	*T. lewisi*	Tcs44	*T. congolense* savannah IL 3000 27/09/00
Tcf2	*T. congolense* forest Komoé/87/CRTA/153	Tcs45	*T. congolense* savannah Samo/82/CRTA/22
Tb3	*T. b. brucei* Nyarafo B.1818	Tcs46	*T. congolense* savannah Samo/28BF
Tb4	*T. b. brucei* Farakoba/81/CRTA/4	Tcs47	*T. congolense* savannah Samo/80/CRTA/20
Tb5	*T. b. brucei* Fol/02/CIRDES/01	Tcs48	*T. congolense* savannah Satiri/87/CRTA/129.1
Tb6	*T. b. brucei* Gaoua/89/CRTA/259	Tcs49	*T. congolense* savannah KRA/98/CIRDES/2344
Tb7	*T. b. brucei* Komoé/87/CRTA/150	Tcs50	*T. congolense* savannah KRA/98/CIRDES/1949
Tv8	*T. vivax* Satiri/87/CRTA/133.3	Tcs51	*T. congolense* savannah KRA/99/CIRDES/1922
Tv9	*T. vivax* Menegué B.1643 du 7/05/96	Tcs52	*T. congolense* savannah Kigoni/ILRAD/776
Tv10	*T. vivax* Satiri/87/CTRA/123.2	Tcs53	*T. congolense* savannah Satiri/CRTA/238.1 B.752
Tv11	*T. vivax* Satiri/87/CRTA/125.1 31.10.02	Tcs54	*T. congolense* savannah Satiri/CRTA/237.2 B.760
Tv12	*T. vivax* Kanfiala (Mali)	Tcs55	*T. congolense* savannah Samo/82/CRTA/31.1
Tv13	*T. vivax* Guyanne	Tcs56	*T. congolense* savannah Satiri/CRTA/197.1 B.705
Tv14	*T. vivax* Lahi/02/CIRDES/02	Tcs57	*T. congolense* savannah TBU/98/CIRDES/155
Tv15	*T. vivax* Satiri/87/CRTA/122.1	Tcs58	*T. congolense* savannah FMA/98/CIRDES/1788
Tv16	*T. vivax* Zaria	Tcs59	*T. congolense* savannah SRI/98/CIRDES/146
Tv17	*T. vivax* Banan/83/CRTA/73	Tcs60	*T. congolense* savannah Bobo/82/CRTA/44.2
Tv18	*T. vivax* Lahi/02/CIRDES/03	Tcs61	*T. congolense* savannah Samo/89/CRTA/267
Tv19	*T. vivax* Lahi/02/CIRDES/01	Tcs62	*T. congolense* savannah Folonzo/02/CIRDES/01.2
Tv20	*T. vivax* Satiri/87/CRTA/120.4	Tcs63	*T. congolense* savannah Samo/CRTA/37.3
Tv21	*T. vivax* Lobi/CRTA/282.1	Tcs64	*T. congolense* savannah Fol/02/CIRDES/01
Tv22	*T. vivax* Nyarafo/96/CIRDES/1	Tcs65	*T. congolense* savannah Banan/83/CRTA/72
Tv23	*T. vivax* Gaoua/89/CRTA/282	Tcs66	*T. congolense* savannah Nyarafo BMI 24/05/96
Tv24	*T. vivax* Fol/02/CIRDES/01		*T. congolense* IL3946
Tv25	*T. vivax* Karankasso/CRTA/58		*T. congolense* K30cl1
Tv26	*T. vivax* Satiri/86/CRTA/100		*T. congolense* IL C66
Tv27	*T. vivax* Noumousso/79/2		*T. congolense* IL3666
Tv28	*T. vivax* Matourkou/97/CRTA/1		*T. congolense* IL3304
Tv29	*T. vivax* Noronin/80/CRTA/16		*T. congolense* Kilifi K60 Kenya
Tv30	*T. vivax* Sarfalao/80/CRTA/1		*T. congolense* forest Komoé/87/CRTA
Tv31	*T. vivax* Bossora/88/CRTA/249		*T. vivax* IL3638
Tv32	*T. vivax* Folonzo/02/CIRDES/02		*T. vivax* IL3671
Tv33	*T. vivax* Sarfalao/83/CRTA/54		*T. vivax* IL2186
Tv34	*T. vivax* Bobo/88/CRTA/243		*T. vivax* IL3245
Tck35	*T. congolense* Kilifi K60		*T. vivax* IL3769
Tcs36	*T. congolense* savannah Samo/82/CRTA/51		*T. congolense* IL3000
Tcs37	*T. congolense* savannah Samo/82/CRTA/53		*T. simiae*
Tcs38	*T. congolense* savannah Samo/83/CRTA/66	Tv2160	*T. vivax* 2160
Tcs39	*T. congolense* savannah Samo/82/CRTA/33		*T. vivax*
Tcs40	*T. congolense* savannah Samo/89/CRTA/253		*T. vivax* ILdat 1.2
Tcs41	*T. congolense* savannah Bobo/40/CRTA/1	TbILtat	*T. b. brucei* ILtat 1.1
Tcs42	*T. congolense* savannah IL 3000	TbCP23	*T. b. brucei*
Tcs43	*T. congolense* savannah Serengetti/71/STIB/212		*T. equiperdum*

2.3   Agarose gel electrophoresis of DNA

Agarose is a high molecular weight polymer extracted from seaweed. It is composed of alternating galactose and 3,6-anhydrogalactose monosaccharides residues. Agarose DNA electrophoresis is the main technique used to fractionate nucleic acids based on size (Sambrook *et al.,* 1989). When an electric current is applied to an agarose gel in the presence of buffer ions, DNA fragments move through the gel towards the anode as DNA is negatively charged. Due to the pore size of the agarose gel, small linear DNA fragments move more rapidly than larger ones. Different concentrations of agarose will allow the optimal resolution of fragments of different size ranges. The DNA sample is contained in a loading buffer and placed in wells cut into the agarose gel. The loading dye contains (i) glycerol that makes the DNA sample dense, allowing it to settle at the bottom of the well and (ii) tracking dye that does not interact with the DNA, and shows the progress of the electrophoretic separation as the dye migrates more rapidly than the DNA. Agarose gel electrophoresis was used to verify amplification of DNA by PCR (Section 2.4), RT-PCR (Section 2.5) and various cloning steps.

2.3.1   Materials

50× Tris-acetate-EDTA (TAE) buffer. Tris (242.0 g), glacial acetic acid (57.1 ml) and EDTA (18.612 g) were dissolved in 800 ml of dH_2O. The pH was adjusted to 8.0 with acetic acid and the volume was made up to 1000 ml with dH_2O. The solution was autoclaved (121ºC, 30 min, RT) and allowed to cool before use.

1× TAE working buffer solution. 50× TAE buffer (20 ml) was diluted to 1000 ml in dH_2O and ethidium bromide was added to a final concentration of 0.5 µg/ml.

1% (w/v) Agarose. Agarose (1.0 g) was suspended in 100 ml of 1× TAE buffer and heated in a microwave oven until the agarose was completely melted. The volume was adjusted with dH_2O to 100 ml after minimal volume loss due to evaporation.

0.8% (w/v) Agarose. Agarose (0.8 g) was suspended in 100 ml of 1× TAE buffer and heated in a microwave oven until the agarose was completely melted. The volume was adjusted with dH_2O to 100 ml after minimal volume loss due to evaporation.

2.3.2   Procedure

The casting tray was assembled and positioned on a level surface. The sample comb was placed one cm from one end of the casting tray. The 1% (w/v) agarose gel was cooled to ~ 55°C and poured carefully into the assembled casting tray. The agarose was allowed to set at RT, the comb carefully removed, and the casting tray placed into the electrode chamber filled

with $1\times$ TAE buffer to submerge the gel. DNA samples were mixed 1:5 with $6\times$ loading dye and loaded into the wells alongside molecular weight markers. Electrophoresis was carried out for 1-2 h at 100 V. The gel was visualised by trans-illumination under UV light and the electrophoregram documented using the Versa Doc™ gel documentation system (Bio Rad).

2.4   Polymerase chain reaction

Polymerase chain reaction (PCR) is a technique for amplifying DNA sequences *in vitro* (Saiki *et al.,* 1985; Mullis and Faloona, 1987). A pair of oligonucleotide primers is designed to flank the region targeted for amplification. The primers anneal to the opposite DNA strands and face inwards so that DNA synthesis proceeds across the central region. Typical PCR amplification reaction components include 5-500 ng of template DNA, a primer pair (forward and reverse primer), 100-200 µM deoxyribonucleotide triphosphate (dNTPs) mix, 1.5 units of thermostable DNA polymerase, PCR reaction buffer and 1.0-4.0 mM $MgCl_2$. There are three major steps in PCR namely: denaturation, annealing and polymerisation. In the denaturation step, the target double stranded-DNA is separated into two strands by heating at 94-95°C. Oligonucleotide primers anneal to the resulting single-stranded template at a lower temperature ranging from 35-65°C for 30-60 s. During polymerisation, the temperatures is raised to 68-74°C for 1-2 min for synthesis of new DNA across the target region, as most thermostable DNA polymerases have optimal activity at these temperatures. Each group of three reactions makes a PCR cycle and the amount of DNA product increases exponentially with every cycle; 20–40 cycles are commonly used.

Taq DNA polymerase is the most commonly used thermostable DNA polymerase. It is isolated from *Thermus aquaticus* (Saiki *et al.,* 1988). The enzyme has 5'-3' exonuclease activity but lacks 3'-5' exonuclease activity. The enzyme yields PCR products with unpaired 3'-deoxyadenine nucleotide (3'-A) overhangs (Clark, 1988; Hu, 1993) which can be conveniently cloned into a plasmid containing 5'-deoxythymidine nucleotide (5'-T) overhangs (T-vectors). The PCR product and the T-vector are joined by a ligase. The success of any PCR reaction depends on the optimal concentration of reaction components. At low magnesium concentration *Taq* DNA polymerase is inactive while excess magnesium leads to non-specific amplifications (Williams, 1989). The use of annealing temperatures slightly higher than the melting temperature (T_m) of the primers will reduce non-specific annealing of primers, resulting in decreased amounts of unwanted products. Annealing temperatures lower than the melting temperatures of the primers may result in higher yields of the PCR product

since the primers anneal more efficiently, but at the risk of losing its specificity. The length of the PCR product determines the extension time, generally one min of extension is sufficient to synthesise a 1 kb long DNA fragment.

2.4.1 Materials

10× PCR buffer II, (provided in the AmpliTaq[®] DNA Polymerase with GeneAmp[®] PCR kit, Table 2.1)

25 mM MgCl$_2$, (provided in the AmpliTaq[®] DNA Polymerase with GeneAmp[®] PCR kit, Table 2.1)

5 U/µl Ampli*Taq*[®] DNA polymerase, (provided in the AmpliTaq[®] DNA Polymerase with GeneAmp[®] PCR kit, Table 2.1)

Oligonucleotide primers as described in the respective Chapters

DNA templates as detailed in Table 2.3

2.4.2 Method

All pipette tips and microfuge tubes were sterilised by autoclaving [121ºC, 30 min, RT]. The PCR protocol was carried out on an area of the laboratory bench designated for PCR to avoid contamination.

A master mix of the PCR reagents was prepared as shown in Table 2.4. Master mix (48 µl) was aliquoted into individual 0.2 ml nuclease-free thin wall PCR tubes to which 2 µl (5-500 ng) of genomic DNA template was added. The PCR mix was subjected to PCR cycling conditions: denaturation at 94°C for 5 min and 94°C for 45 s, annealing at 50-60°C for 45 s, polymerisation at 72°C for 1-2 min (30 cycles) and final extension at 72°C for 7 min.

Table 2.4 PCR reaction components

PCR components	Volume 50µl	Final concentration
DNA template	2	5-500 ng
Fw primer	0.2	0.2-1 µM
Rev primer	0.2	0.2-1 µM
10× PCR buffer II	5	1×
25 mM MgCl$_2$	5	2.5 mM
10 mM dNTPs mix	1	200 mM each
Ampli*Taq*[®] DNA polymerase	0.25	1.25 units
Sterile dH$_2$O	36.35	

2.5 Reverse transcriptase-polymerase chain reaction (RT-PCR)

Thermostable DNA polymerase used in basic PCR is limited to use with DNA as template. Due to the importance of targeting RNA, an alternative enzyme is required. For use of thermostable DNA polymerase in amplification of RNA, the RNA must be converted into DNA in a process called reverse transcription catalysed by reverse transcriptase. Alternatively some reverse transcriptases such as avian myloblastosis virus (AMV) or Moloney murine leukaemia virus (M-MLV or MuLV) can be used to produce a DNA copy of the RNA template using random primers, an oligo (dT) primer or a gene-specific primer. Currently it is possible to carry out RT-PCR using a single enzyme such as T*th* or *Tfl* DNA polymerase from *Thermus thermophilus* or *T. flavus* which has a reverse transcriptase activity in the presence of manganese ions and a DNA polymerase activity in the presence of magnesium ions (Myers and Gelfand, 1991).

Access RT-PCR system (Table 2.1) was used for RT-PCR on mRNA purified from *T. congolense* clone IL 3000 parasites. This system facilitates one-tube PCR, whereby AMV reverse transcriptase (AMV RT) catalyses synthesis of the first DNA strand and thermostable *Tfl* DNA polymerase from *Thermus flavus* catalyses the second strand cDNA synthesis and amplification.

2.5.1 Materials

Nuclease free water, (provided in the Access RT-PCR kit, Table 2.1)

AMV/*Tfl* 5× reaction buffer, (provided in the Access RT-PCR, Table 2.1)

10 mM dNTPs mix, 25 mM $MgSO_4$, (provided in the Access RT-PCR, Table 2.1)

AMV Reverse transcriptase (5U/µl), (provided in the Access RT-PCR, Table 2.1)

Tfl DNA polymerase (5U/µl), (provided in the Access RT-PCR, Table 2.1)

Oligonucleotide primers as detailed in the respective Chapter

RNA templates (Section 2.5.2)

2.5.2 Method

All pipette tips and microfuge tubes were sterilised by autoclaving (121ºC, 30 min, RT). Diethylpyrocarbonate (DEPC) was used for cleaning all apparatus and the bench. PCR protocol was carried out at an area of the bench designated for PCR.

Total RNA was isolated from *T. congolense* clone IL 3000 parasites using SV Total RNA isolation kit as described by the manufacturer (Table 2.1) and the mRNA subfraction gel

purified according to RNAid Kit protocol. A master mix of the RT-PCR reagents was prepared as shown in Table 2.5 and 24 µl of the master mix aliquoted into individual 0.2 ml nuclease-free thin wall PCR tubes to which 0.5 µl of both forward and reverse primers were added. The PCR mix was subjected to one cycle for reverse transcription at 45°C for 45 min, 94°C for 2 min, 94°C for 30 s, 60°C for 1 min, 68°C for 2 min (40 cycles) and 68°C for 7 min.

Table 2.5 RT-PCR reaction components

Components	Volume 25 µl	Final concentration
Nuclease free water	15.5	
AMV/*Tfl* 5× reaction buffer	5	1×
10 mM dNTPs mix	0.5	200 mM each
C1, C2 and C3 Fw/Cluster 1-3 Fw primers	0.5	0.2-1 µM
C1, C2 and C3 Rev/Cluster 1-3 Rev primers	0.5	0.2-1 µM
25 mM $MgSO_4$	1	5 mM
AMV Reverse transcriptase (5u/µl)	0.5	2.5 units
Tfl DNA polymerase (5u/µl)	0.5	2.5 units
RNA template	1	~ 100 ng

2.6 Purification of DNA from agarose gels

The purpose of the purification of the PCR fragments was to remove the residual DNA polymerases and reaction buffer, unincorporated nucleotides, primers and non-specific amplification products following PCR and RT-PCR. The removal of these contaminants enhances cloning efficiency in subsequent steps. Like all commercial nucleic acid purification kits, it relies on the binding of DNA excised from the agarose gel on silica matrix.

2.6.1 Materials

Isopropanol [80% (v/v)]. Isopropanol (80 ml) was made up to 100 ml with sterile dH_2O and stored at RT.

Wizard® SV Gel and PCR clean-up system kit (Table 2.1)

2.6.2 Procedure

The Wizard® SV Gel and PCR clean-up system kit (Table 2.1) was used according to the manufacturer's protocol to purify DNA from agarose gel.

2.7   Cloning of PCR products into pGEM®-T Easy vector

T4 DNA ligase catalyses the joining of two DNA stands between 5'- phosphate and 3'- OH groups of adjacent nucleotides to form a phosphodiester bond (Weiss *et al*., 1968). The pGEM® -T Easy vector (Figure 2.1) is engineered for cloning PCR products because it has 3'-T overhangs at both ends. *Taq* DNA polymerase adds single 3'-A overhangs to double stranded DNA during the PCR reaction (Promega pGEM®-T Easy technical manual). The complementarity between the pGEM®-T Easy vector T-overhangs and PCR products 3'–A overhangs improves the efficiency of ligation. The 3'-T overhangs also prevent vector recircularisation and hence enhance ligation efficiency. pGEM®-T Easy is a high copy plasmid containing T7 and SP6 RNA polymerase promoters flanking a multiple cloning site (MCS) with the α-peptide coding region of the enzyme β-galactosidase. Insertional inactivation of the α-peptide allows recombinant clones to be directly identified by white and blue screening in the presence of X-gal and IPTG. The MCS is flanked by recognition sites for the restriction enzymes *Eco*RI, *Bst*ZI and *Not*I, thus providing three restriction enzymes for release of the insert. The pGEM®-T Easy vector also contains the origin of replication of filamentous phage f1 for the preparation of single-stranded DNA (ssDNA) (Promega pGEM®-T Easy technical manual).
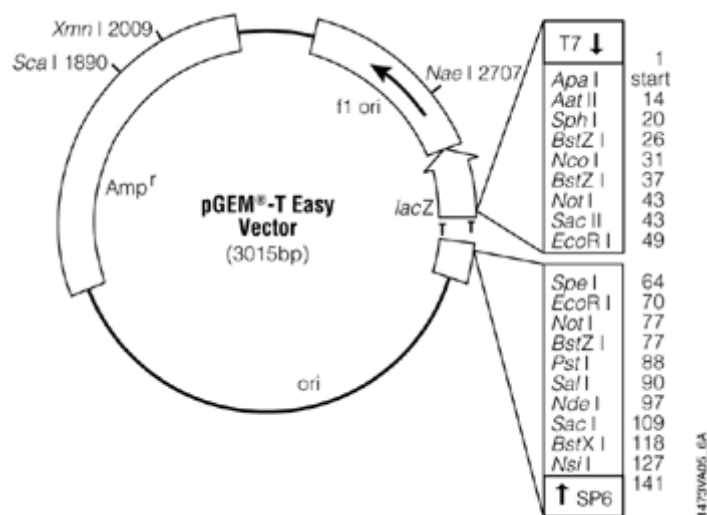
Figure 2.1 Map of pGEM®-T Easy vector showing its multiple cloning sites (www.promega.com accessed on 01.10.06)

2.7.1   Materials
pGEM®-T Easy vector kit [50 ng/μl] (Table 2.1)

2.7.2 Procedure

The ligation reaction mix was prepared by combining 2× Rapid ligation buffers (5 µl), pGEM®-T Easy vector (12.5 ng), T4 DNA ligase (0.5 µl) and insert (40-100 ng) and the volume was adjusted to 10 µl by the addition of dH$_2$O. Ligations were done at RT for 1h.

2.8 Preparation of competent *E. coli* JM 109 cells

Competency of *E. coli* cells is defined as the ability of the cells to take up foreign DNA and it is expressed as cfu (colony forming unit) per µg DNA. There are two methods for transforming bacterial cells: chemical transformation where bacterial cells are treated with divalent ions (Ca^{2+}) to make them competent, and by electroporation whereby the bacterial cells are exposed to an electric field to create pores through which DNA enters (Dagert and Ehrlich, 1979). Both methods of transformation were used in the present study. Competent *E. coli* cells were obtained from Promega Corporation [pGEM®-T Easy Vector System II and Invitrogen Life Technologies (Table 2.1)]. Due to the high cost involved in purchasing chemically competent cells from manufacturers, these were prepared using several methods i.e. The TranformAid™ Bacterial transformation kit from Fermentas life sciences, calcium chloride and a modified TB calcium chloride method (Inoue *et al.,* 1990). Competent cells prepared by TranformAid™ Bacterial transformation kit were very good for transformation, while those using the calcium chloride method were less competent. Although the modified calcium chloride method provided cells with high competence, commercial competent cells were preferred where large numbers of clones were required.

2.8.1 Preparation of competent *E. coli* JM 109 cells using the TransformAid™ Bacterial Transformation kit

2.8.1.1 Materials
Ampicillin [50 mg/ml]. Ampicillin (0.5 g) was dissolved in dH$_2$O (10 ml), filtered through 0.45 µm filters and stored in a 15 ml conical centrifuge tube at 4°C.

Ampicillin [100 mg/ml]. Ampicillin (1.0 g) was dissolved in dH$_2$O (10 ml), filtered through 0.45 µm filters and stored in a 15ml conical centrifuge tube at 4°C.

2× YT medium. Bacto-tryptone (16.0 g), yeast extract (10.0 g) and NaCl (5.0 g) were dissolved by shaking in dH$_2$O (900 ml). The pH was adjusted to 7.0 with NaOH, and made up to 1000 ml with dH$_2$O. Half of the medium was sealed and autoclaved (121ºC, 30 min, RT) for preparation of bacterial cultures. The remaining half was used for preparation of 2× YT-ampicillin agar plates.

2× YT-ampicillin agar plates. Agar (7.5 g) was dissolved in 2× YT media (500 ml) and autoclaved (121ºC, 30 min, RT), cooled to 50°C before adding ampicillin to a final concentration of 100 µg/ml. Agar (30-35 ml) was in poured into 85 mm Petri dishes and allowed to solidify.

Glycerol stocks of competent *E. coli* JM 109 cells

Sterile glycerol [10% (v/v)] Glycerol (10 ml) was made up to 100 ml with sterile dH_2O, autoclaved (121ºC, 30 min, RT) and kept at 4°C after cooling.

2.8.1.2 Procedure

Competent *E. coli* JM 109 cells were prepared using TranformAid™ Bacterial transformation kit according to the manufacturer's specification.

2.8.2 Preparation of competent *E. coli* JM 109 cells using the calcium chloride method

2.8.2.1 Materials
Glycerol stocks of *E. coli* JM 109 cells

0.1 M calcium chloride $CaCl_2.2H_2O$ (14.7 g) was dissolved in 900 ml of dH_2O, the volume adjusted to 1000 ml and autoclaved (121ºC, 30 min, RT).

0.1 M magnesium chloride. $MgCl_2.6H_2O$ (20.3 g) was dissolved in 900 ml of dH_2O, the volumes adjusted to 1000 ml and autoclaved (121ºC, 30 min, RT).

2.8.2.2 Procedure
A fresh colony of *E. coli* JM 109 cells was inoculated into 5 ml of 2× YT and incubated overnight in a shaker at 37°C. Overnight *E. coli* JM 109 cell culture (1 ml) was placed in 100 ml of fresh 2× YT in a 500 ml conical flask, and incubated at 37°C in shaker until an OD_{600}= 0.4-0.5 was attained. The cells were centrifuged (6000×g, 5 min, 4°C), the medium drained off, the pellet resuspended in 50 ml of 0.1 M ice-cold $MgCl_2$ and incubated on ice for 30 min. The cells were centrifuged as before, the $MgCl_2$ drained off and the pellet resuspended in 10 ml of ice cold 0.1 M $CaCl_2$. The cells were stored at -70°C after adding sterile 50% (v/v) glycerol to a final concentration of 20% (v/v) or immediately used for transformation.

2.8.3 Preparation of competent *E. coli* JM 109 cells with modified TB

2.8.3.1 Materials
TB ($CaCl_2$) solution: PIPES (3.021 g), $CaCl_2.2H_2O$ (2.205 g) and KCl (18.637 g) were dissolved in dH_2O (900 ml) and the pH was adjusted to 6.7 with KOH. $MnCl_2$ (10.885 g) was

dissolved in the solution, the volume was made up to 1000 ml of dH_2O and filtered through a 0.45 μM membrane. The solution was stored at 4°C.

S.O.C medium: Bacto-tryptone (20 g), Bacto-yeast extract (5 g) and NaCl (0.5 g) were dissolved in dH_2O (900 ml). 1 M $MgSO_4$ (10 ml), 1 M $MgCl_2$ (2.5 ml) were added and the pH adjusted to 7 with NaOH and the volume made up to 990 ml with dH_2O. The medium was autoclaved (121ºC, 30 min, RT) and filtered. Glucose (2 M, 10 ml) was added after cooling.

S.O.C-ampicillin agar plates. Agar (7.5 g) was dissolved in S.O.C medium (500 ml) and autoclaved (121ºC, 30 min, RT), cooled to 50°C before adding glucose (2 M, 5 ml) and ampicillin to a final concentration of 100 μg/ml. Agar (30-35 ml) was poured into 85 mm Petri dishes and allowed to solidify.

2.8.3.2 Procedure
An overnight culture of *E. coli* JM 109 cells (100 μl) was inoculated into fresh S.O.C medium (50 ml) and incubated at 37°C for 1 h. The JM 109 *E. coli* culture was grown in a 37°C shaker until the cells attained an OD_{600} of 0.5; the culture (20 ml) was transferred into sterile centrifuge tubes and centrifuged (6000×*g*, 2 min, 4°C). The medium was drained off, cells resuspended in cold TB ($CaCl_2$) solution (10 ml) and incubated on ice for 25 min. This procedure was repeated twice and the final cell pellet resuspended in cold TB $CaCl_2$ solution (2 ml). The cells were stored at -70°C after adding sterile 50% (v/v) glycerol to a final concentration of 20% (v/v) or immediately used for transformation.

2.8.4   Preparation of *E. coli* JM 109 cells for electroporation

2.8.4.1 Materials
Glycerol stocks of competent *E. coli* JM 109 cells
Sterile glycerol [10% (v/v)] (Section 2.8.1.1)

2.8.4.2 Procedure
A fresh colony of *E. coli* JM 109 cells was inoculated into 2×YT medium (Section 2.8.1.1) and grown in a shaker overnight at 37°C. The overnight cell culture (50 ml) was diluted into fresh 2× YT medium (500 ml) in a 1000 ml flask, the cells grown for 2-3 h until $OD_{600} = 0.6$. The cells were transferred into sterile 50 ml conical centrifuge tubes and harvested by centrifugation (5000×*g*, 10 min, 4°C). The medium was poured off, the cell pellet washed in ice-cold sterile water (25 ml), and centrifuged as before. The water was drained carefully and cells washed again in ice-cold sterile water (12.5 ml) and centrifuged as before. The water

was drained off again and the cells were resuspended in ice-cold 10% (v/v) glycerol (6.25 ml). After centrifugation as before, the glycerol was poured off and the cells resuspended in ice-cold 10% (v/v) glycerol (5 ml). Aliquots (100 µl) were used immediately or frozen at -70°C.

2.9 Transformation

Various methods for preparing competent cells were described (Section 2.8), in this section transformation protocols for each of the methods are given.

2.9.1 Transformation of pGEM®-T Easy vector containing inserts into competent *E. coli* JM 109 cells prepared by TransformAid™ Bacterial Transformation kit

2.9.1.1 Materials

TransformAid™ C-medium (provided with the TransformAid™ Bacterial Transformation kit, Table 2.1)

T-Solution A and T-Solution B (provided with the TransformAid™ Bacterial Transformation kit, Table 2.1)

Isopropyl-β-D-thiogalactopyranoside (IPTG) [100 mM]

IPTG (1.2 g) was dissolved in 40 ml of dH$_2$O, and made up to a final volume of 50 ml with dH$_2$O. The IPTG was sterilized by filtration through a 0.22 µm disposable filter.

5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) [50 mg/ml]

X-gal (100 mg) was dissolved in 2 N, N'- dimethylformamide (2 ml).

Competent *E. coli* JM 109 cells (Section 2.8.2.2)

Ampicillin [50 mg/ml] (Section 2.8.1.1)

2× YT-ampicillin agar plates (Section 2.8.1.1)

2.9.1.2 Procedure

Competent *E. coli* JM 109 cells were transformed by ligation mixtures according to the manufacturer's specification. The transformed cells were plated on 2× YT agar plates containing 100 µg/ml ampicillin, 50 mg/ml X-gal (10 µl) and 100 mM IPTG (20 µl). The plates were incubated overnight at 37°C.

2.9.2 Transformation of cloning vectors containing inserts into *E. coli* JM 109 cells prepared by calcium chloride method

2.9.2.1 Materials
Competent *E. coli* JM 109 cells (Section 2.8.2.2)

2× YT media (Section 2.8.1.1)

Ampicillin [50 mg/ml] (Section 2.8.1.1)

X-gal [50 mg/ml] (Section 2.9.1.1)

100 mM IPTG (Section 2.9.1.1)

2× YT-ampicillin agar plates (Section 2.8.1.1)

2.9.2.2 Procedure
Ligation mix (2 µl) was added to thawed competent cells (100 µl) in a 15 ml sterile conical centrifuge tube placed on ice and incubated for 30 min, this was followed by incubation at 42°C for 2 min. 2× YT medium (900 ml) was added and the sample incubated in a shaker at 37°C for 1 h. The transformed cells were plated on 2× YT agar plates containing 100 µg/ml ampicillin, 50 mg/ml X-gal (10 µl) and 100 mM IPTG (20 µl). The plates were incubated overnight at 37°C.

2.9.3 Transformation of cloning vectors containing inserts into competent *E. coli* JM 109 cells treated with modified TB (calcium chloride)

2.9.3.1 Materials
Competent *E. coli* JM 109 cells (Section 2.8.3.2)

Ampicillin [50 mg/ml] (Section 2.8.1.1)

X-gal [50 mg/ml] (Section 2.9.1.1)

100 mM IPTG (Section 2.9.1.1)

S.O.C medium (Section 2.8.3.1)

DMSO (Section 2.2)

2.9.3.2 Procedure
Ligation mix (2-4 µl) was added to competent cells (100 µl), DMSO (1 µl) added and the cells were incubated on ice for 30 min. The cells were placed in a water bath at 42°C for 90 s and incubated on ice for 2 min. S.O.C medium (400 ml) was added and the cells were incubated at 37°C with shaking for 45 min. Transformed cells (100 µl) were plated on pre-

warmed S.O.C agar plates containing ampicillin (100 μg/ml), 50 μg/ml X-gal and 100 μM IPTG and incubated overnight at 37°C.

2.9.4. Electroporation of electro-competent *E. coli* JM 109 cells

2.9.4.1 Materials
Electro-competent *E. coli* JM 109 cells (Section 2.8.4.2)

2× YT medium (Section 2.8.1.1)

Ampicillin [50 mg/ml] (Section 2.8.1.1)

X-gal [50 mg/ml] (Section 2.9.1.1)

100 mM IPTG (Section 2.9.1.1)

2× YT-ampicillin plates (Section 2.8.1.1)

2.9.4.2 Procedure
Cells (100 μl) were thawed on ice and ligation mix (2 μl) added to the cells. The cells were incubated on ice for 30 min, transferred into cold 0.2 cm electroporation cuvettes (BioRad) on ice, and were pulsed at 25 μF, 200 Ω and 2.5 kV for 4.5-4.8 ms. Sterile 2× YT medium (900 μl) was added to the cells after the pulse and the cell suspension transferred into 15 ml conical centrifuge tubes. The cells were grown at 37°C with agitation for 1 h in a shaker at 225 rpm and the transformed cells (200 μl) were spread on 2× YT agar plates containing ampicillin (100 μg/ml), 50 mg/ml X-gal (10 μl) and 100 mM IPTG (20 μl) and plates were incubated overnight at 37°C.

2.9.5 Transformation of pGEM®-T Easy vector containing PCR products

2.9.5.1 Materials
Competent JM 109 *E. coli* cells >10^8 Cfu/μg (Table 2.1)

2× YT media (Section 2.8.1.1)

Ampicillin [50 mg/ml] (Section 2.8.1.1)

X-gal [50 mg/ml] (Section 2.9.1.1)

100 mM IPTG (Section 2.9.1.1)

2× YT-ampicillin agar plates (Section 2.8.1.1)

2.9.5.2 Procedure
Ligation mix (2 μl) was added to thawed competent cells (20 μl) in a 15 ml sterile conical centrifuge tube placed on ice and incubated for 20 min, followed firstly by incubation at 42°C for 45 s and secondly on ice for 2 min. 2× YT media (200 ml) was added and the

sample incubated in a shaker at 37°C for 1 h. The transformed cells were plated on 2× YT agar plates containing 100 µg/ml ampicillin, 50 mg/ml X-gal (10 µl) and 100 mM IPTG (20 µl). The plates were incubated overnight at 37°C.

2.10 Screening of recombinant plasmids by colony PCR

Colonies carrying the cloned fragments were identified by PCR using universal primers SP6 and T7 (Section 2.7)

2.10.1 Materials

SP6 primer: TATTTAGGTGACACTATAG, T_m (50ºC)

T7 primer: TAATACGACTCACTATAGGG, T_m (56 ºC)

Sterile tooth pick

PCR reagents (Table 2.4)

2.10.2 Procedure

A sterile tooth pick was used to transfer part of a white colony into 0.2 ml PCR tubes containing PCR master mix (25 µl). The PCR tubes were flicked at the bottom to mix the cells with PCR reagents and centrifuged briefly. The tubes were subjected to PCR cycling conditions: 94°C for 5 min (initial denaturation), denaturation at 94°C for 45 s, annealing at 50°C for 45 s, polymerisation 72°C for 1-2 min for 25 cycles and final extension at 72°C for 7 min.

2.11 Purification of plasmid DNA (Minipreps)

Plasmid DNA was obtained by minipreparation of plasmids (minipreps). Wizard® *Plus* SV Miniprep DNA purification system exploits the alkaline/SDS-denaturation method for purifying plasmid DNA (Birnboim and Doly, 1979). The SDS-denaturation method is based on the differences in denaturation and renaturation characteristics between plasmid DNA and chromosomal DNA. Denaturation of both plasmid and chromosomal DNA occurs under alkaline conditions. Upon addition of neutralisation solution, the chromosomal DNA rehybridises in an intra-strand manner, hence forming an insoluble aggregate that precipitates, whereas plasmid DNA undergoes interstrand rehybridisation and remains in solution. Potassium forms an insoluble salt with SDS, which precipitates proteins and complexes together with chromosomal DNA, leaving the plasmid DNA in the clear lysate. The chromosomal and plasmid DNA are then separated by centrifugation; any RNA present is digested by RNases present in the resuspension solution. The clear lysate formed is then transferred onto a column containing silica where the DNA binds in the presence of a high

concentration of guanidine hydrochloride (Marko *et al*., 1982; Boom *et al*., 1990). Two ethanol washes remove the salts and the DNA is eluted in a low ionic strength solution such as TE buffer or nuclease free water.

2.11.1 Materials
2× YT media (Section 2.8.1.1)

Ampicillin [50 mg/ml] (Section 2.8.1.1)

Wizard® *Plus* SV Minipreps DNA purification kit (Table 2.1)

2.11.2 Procedure
A single colony from a 2× YT agar plate was inoculated into the 2× YT medium (5 ml) containing ampicillin (100 µg/ml) and incubated overnight (12-16 h) at 37°C in a shaking incubator. The bacterial culture was centrifuged (5000×g, 5 min, RT) in a table top centrifuge, the supernatant was poured off and the pellet resuspend in resuspension solution (provided in the Wizard® *Plus* SV Minipreps DNA purification kit) and transferred into 1.5 ml microfuge tubes. The plasmid DNA was purified using Wizard® *Plus* SV Minipreps DNA purification kit according to the manufacturer's instructions.

2.12   DNA purification using Wizard® DNA Clean-Up System
Good quality DNA is required in all DNA manipulation steps. Wizard® DNA Clean-Up System purifies DNA from salts, enzymes and also concentrates DNA. Purification using this kit relies on the same principle as the Wizard SV Gel and PCR clean-up system as described in Section 2.6. The chaotropic buffer contains silica (resin) to which DNA binds. A sequence of ethanol washes and centrifugation steps removes contaminants from the DNA and the DNA is then eluted from the silica mini-columns with a low salt concentration solution such as TE or nuclease free water.

2.12.1 Materials
Wizard® DNA Clean-Up System (Table 2.1)

Isopropanol [80% v/v] (Section 2.6.1)

2.12.2 Procedure
The DNA was purified using Wizard® DNA Clean-Up System according to the manufacturer's specifications.

2.13   DNA sequencing
DNA fragments are sequenced to determine the order of the nucleotides in a given fragment. There are two major methods of DNA sequencing, i.e. the chain termination method (Sanger

et al., 1977) and chemical sequencing (Maxam and Gilbert, 1977). Chain termination sequencing also called Sanger sequencing, is the most widely used method and was used in the present study. This method is based on the replacement of the 3'-OH group by hydrogen to result in a 2', 3'-dideoxynucleotide that cannot form a phosphodiester bond with the next nucleotide and DNA chain elongation is thus terminated. Oligonucleotide primers are used to initiate sequencing, and like in normal PCR reactions the chain is elongated by DNA polymerase in the presence of dNTPs and a low concentration of four dideoxynucleotides (ddATP, ddGTP, ddCTP and ddTTP). DNA polymerase elongates the chain until a dideoxynucleotide is incorporated into the growing strand. The different-sized fragments are separated on a polyacrylamide gel. Originally, the sequencing process involved 4 independent reactions, each with one terminating ddNTP. Currently dideoxynucleotides are labelled with different fluorescent dyes with different emission wavelengths that allow reading of the nucleotide sequence with a fluorescent scanner. All sequencing was carried out at the International Livestock Research Institute (ILRI), Nairobi, Kenya.

2.14   Bioinformatics

A variety of bioinformatics tools were used in the present study: sequence alignment, use of homology (BLAST) programmes for identification of sequence, nucleotide translation into protein and a study of the restriction pattern of congopain-like cysteine proteases. To design the primers used in the present study, alignment of sequences was done by Sequencher 4.5 (Gene Codes Corporation) and ClustalW (www.ebi.ac.uk/ClustalW). These two programmes were also used to align sequences obtained from the study and establish their relatedness. Phylogenetic trees were also constructed using ClustalW. NCBI BLAST (www.ncbi.nlm.nih.gov/BLAST) (Thompson *et al*., 1994) was used to confirm that cloned DNA sequences were part of cysteine protease genes. GeneDb (www.genedb.org) was also used to blast sequences and compare cloned DNA sequences with existing sequences in this database. ExPASy proteomics tools (www.expasy.org/tools) were used to translate nucleotide sequences into predicted amino acid sequences. NEBcutter V2.0 (www.tools.neb.com/NEBcutter2) was used to determine restriction sites in congopain-like cysteine protease sequences.

CHAPTER 3

CONSTRUCTION OF A CONGOPAIN-LIKE CYSTEINE PROTEASE GENE UNITS LIBRARY

3.1   Introduction

Congopain is the major cysteine protease in *T. congolense* infected-cattle (Authié *et al*., 1992) and has been shown to contribute to the pathology of trypanosomosis (Authié *et al*., 1993, 2001; Authié, 1994). Cattle immunized with congopain develop high levels of anti-congopain IgG which inhibit its activity (Authié *et al*., 2001). Cysteine protease inhibitors have been suggested as potential chemotherapeutic agents in parasitic diseases (Sajid and McKerrow, 2002). Development of an anti-disease vaccine and a chemotherapeutic agent based on congopain-like cysteine proteases requires an understanding of their genetic diversity. Therefore, the variability within these families of genes was studied using different approaches. In this chapter, one approach is reported, by constructing and screening a congopain 2 kbp gene unit library from cosmid clones containing various length of the congopain locus.

In a previous study, a cosmid library of *T. congolense* IL 3000 genomic DNA was constructed, screened using the CP2 open reading frame as a probe and 15 positive clones were selected (Dr. Boulangé, manuscript in preparation). The clones were digested with *Pst*I and the released 2 kb gene units cloned into a pGEM®-3Zf(+) vector. Approximately 120 clones were selected and screened by sequencing (24 clones) and PCR (96 clones) to establish the variability within the congopain-like cysteine protease genes present in the locus.

3.2   Materials and Methods

3.2.1   Digestion of CP cosmid clones with *Pst*I
The location of *Pst*I restriction site in the congopain gene is shown in Appendix A.

3.2.1.1 Materials
*Pst*I (Table 2.2)

10× buffer H (provided together with Pst*I*, Table 2.2)

Cosmid clones (Provided by Dr. Boulangé)

3.2.1.2 Procedure

For each of the 15 CP cosmid clones, the digestion reaction consisted of CP cosmid DNA (11 µl), 10× buffer (1.5 µl), *Pst*I (0.5 µl) and made up to 15 µl with dH_2O. The digest was carried out at 37°C for 2 h and the fragments separated on a 0.6% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). For each of the 15 clones the 2 kb fragment was excised from the gel and purified as a pool using Wizard® PCR Preps DNA Purification kit according to the manufacturer's protocol.

3.2.2 Digestion of pGEM®-3Zf(+) cloning vector with *Pst*I

A map of the pGEM®-3Zf(+) cloning vector showing the multiple cloning sites is given in Figure 3.1.

3.2.2.1 Materials

pGEM®-3Zf(+) cloning vector (Promega)

*Pst*I (Table 2.2)

10× buffer H (provided together with Pst*I*, Table 2.2)

3.2.2.2 Procedure

The pGEM®-3Zf(+) cloning vector (1.2 µg) was digested with *Pst*I (1 µl), in 1µl (10× buffer H) and made up to 10 µl with dH_2O. The reaction was incubated at 37°C for 30 min.

3.2.3 Dephosphorylation of digested pGEM®-3Zf(+) cloning vector

3.2.3.1 Materials

Pst*I*-digested pGEM®-3Zf(+) (Section 3.2.3)

Shrimp Alkaline phosphatase (SAP) (Table 2.2)

10× dephosphorylation buffer (provided together with SAP, Table 2.2)

3.2.3.2 Procedure

*Pst*I-digested pGEM®-3Zf(+) (1.0 µg) was dephosphorylated by adding 10× dephosphorylation buffer (1.3 µl), SAP (1.0 µl) and made up to 13 µl with dH_2O. The reaction was done at 37°C for 1 h, the SAP was inactivated in a water bath at 65°C for 30 min. Digested and dephosphorylated pGEM®-3Zf(+) was purified using Wizard® DNA Clean-Up System as described by the manufacturer.
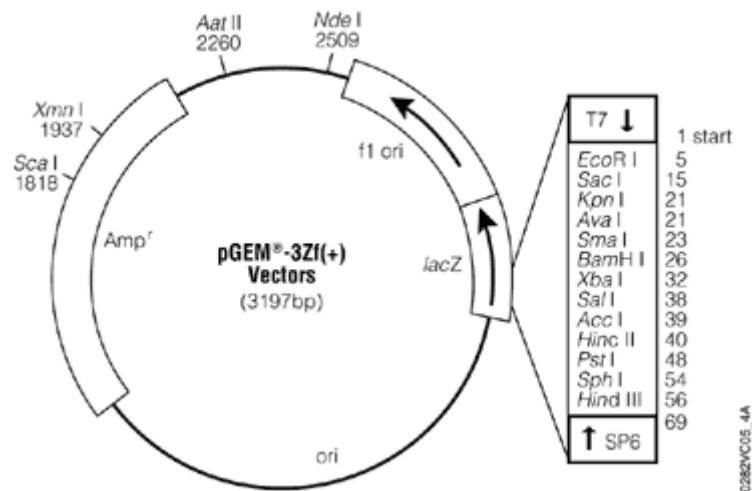
Figure 3.1 Map of pGEM®-3Zf(+) cloning vector (www.promega.com accession date 19th November 2006).

3.2.4 Cloning of the 2 kb CP cosmid *Pst*I digested fragment into *Pst*I digested and dephosphorylated pGEM®-3Zf(+) cloning vector

3.2.4.1 Materials

*Pst*I-digested and dephosphorylated pGEM®-3Zf(+) (Section 3.2.4.2)

2 kb fragment from CP cosmid library (Section 3.2.1.2)

T4 DNA ligase (Table 2.2)

10× ligation buffer (provided together with T4 DNA ligase, Table 2.2)

3.2.4.2 Procedure

To the *Pst*I-digested and dephosphorylated pGEM®-3Zf(+) cloning vector (0.2µg), 2 kb CP cosmid fragment (0.7µg), 10× ligation buffer 1 µl and T4 DNA ligase 1.0 µl were added and ligation carried out overnight at 16°C.

3.2.5 Transformation of ligated pGEM®-3Zf(+) cloning vector/2 kb CP cosmid fragment into competent *E. coli* JM 109 cells

3.2.5.1 Materials

Competent *E. coli* JM 109 cells (Table 2.1)

2× YT medium (Section 2.8.1.1)

2× YT-ampicillin agar plates (Section 2.8.1.1)

3.2.5.2 Procedure

Competent *E. coli* JM 109 cells (100 µl) were transformed by adding ligation mixture (3 µl) and the reaction conducted as per the manufacturer's protocol. The transformed cells were plated on 2× YT agar plates containing 100 µg/ml ampicillin, 50 µg/ml X-gal and 100 µM IPTG. The plates were incubated overnight at 37°C. The recombinant colonies from 2 × YT plates were inoculated in 2× YT medium (5 ml) containing ampicillin (100 µg/ml) and incubated overnight (12-16 h) at 37°C in a shaking incubator. The bacterial culture was centrifuged (6000×g, 5 min, RT) in a table top centrifuge, the supernatant poured off, the pellet resuspended in resuspension solution (provided in the Wizard® *Plus* SV Minipreps DNA purification kit, Table 2.1) and transferred into 1.5 ml microfuge tubes. Plasmid DNA was isolated using Wizard® *Plus* SV Minipreps DNA purification kit (Table 2.1) according to the manufacturer's protocol.

3.2.6   *Pst*I digestion of recombinant pGEM®-3Zf(+)/2 kb clones

3.2.6.1 Materials
Recombinant pGEM®-3Zf(+)/2 kb clones (Section 3.2.5.2)

*Pst*I (Table 2.2)

10× buffer H (provided together with Pst*I*, Table 2.2)

3.2.6.2 Procedure
Recombinant pGEM®-3Zf(+)/2 kb clones 1-30 were digested with *Pst*I in a reaction consisting of recombinant plasmid DNA (5.0 µl), 10× buffer (1 µl), *Pst*I (0.5µl) and made up to 10 µl with dH$_2$O. The digestion was carried out for 1 h at 37°C and separated on a 0.8% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2).

3.2.7   Design of primers for screening the 2 kb CP cosmid library
Preliminary results of sequencing showed that the genes clustered in three groups, named CP1, CP2 and CP3, as they respectively most closely resembled the 3 genes present in the database. In order to establish the proportions of CP1, CP2 and CP3 in the 2 kb CP cosmid library, six different pairs of primers were designed, two sets for each family. The first set, Cluster-C1, -C2, -C3 (Fw and Rev), amplify the same-sized product of 250 bp. The second

set has a single forward primer (Forward-all-CP) common to all CPs and 3 distinct reverse primers (Rev-C1,-C2,-C3), located at different position on congopain, generating different-sized products. It was originally designed to be used as a mix of four primers in a single PCR reaction.

The forward primers Cluster-C1, -C2, -C3, -Fw target the sequence coding for the amino acids around the active cysteine residue (or serine in the case of CP3, see the results section of this chapter), while Cluster-C1, -C2, -C3, -Rv target a sequence situated 250 bp downstream, in a region of less conservation at DNA level (Table 3.1).

The other set of primers was composed of a common forward primer named Forward-all-CP targeting the sequence coding for amino acid residues EAFRF, a region in the propeptide that is conserved in cysteine proteases of the cathepsin L-like subfamily (Sajid and McKerrow, 2002). Rev-C1 primer targeted the genes coding for a region around the histidine residue of the catalytic domain which is unique to CP1 (LDHDVLL), 700 bp downstream of Forward-all-CP primer, Rev-C2 primer targeted an sequence unique to CP2 within the catalytic domain, 415 bp downstream, and Rev-C3 targeted a region at the end of the catalytic domain and unique to a group of CP3-like cysteine proteases, 840 bp downstream (Table 3.1).

Table 3.1 CP cluster and other primers used for screening 2 kb CP cosmid clones

Name	Sequence (5'-3')	Position as from initia ATG	T_m (°C)
Cluster-C1-Fw	AAGGACCAGGGACAATGCGGT	424-444	59
Cluster-C1-Rev	GGCATTTTCCCCCCTTTTGAG	669-650	58
Cluster-C2-Fw	AAGGACCAGGGGCAATGCGGC	424-444	66
Cluster-C2-Rev	GGCACGTTCCCCCCTCCTGAT	669-650	63
Cluster-C3-Fw	GAGAGGTTATGCGACTCTTCG	430-450	51
Cluster-C3-Rev	ACGCGGCACGTCCCCATCTGT	675-665	65
Forward-all-CP	GGAGGCATTCCGTTTCCG	169-186	58
Rev-C1	GAGGAGCACATCATGATCCAGTC	873-851	48
Rev-C2	ACGCGTCGTCCATGAGACCACCTT	585-562	58
Rev-C3-(A4)	GGAGCTCACGTAATCTTTCATAC	1008-986	50
Rev-new-830	ACGGCAATCGCGACGGGCCC	830-810	70

3.2.8 Screening of 2 kb CP cosmid clones by PCR

In addition to the 24 clones sequenced, the remaining 96 were screened by PCR using two separate sets of primers (Section 3.2.7) designed from congopain-like cysteine protease genes based on either CP1, CP2 or CP3 DNA sequences (Table 3.1). These primers were used to establish the proportions of CP1, CP2 and CP3 related clones from the 2 kbp gene unit library, and compared with results obtained by sequencing.

3.2.8.1 Materials
Glycerol stocks of pGEM®-3Zf(+)/2 kb clones (Section 3.2.5.2)
Cluster primers-C1, -C2 and -C3 (Fw and Rev) (Table 3.1)
Forward-all-CP and Rev-C1, Rev- C2 and Rev-C3 primers (Table 3.1)
PCR reagents (Table 2.4)

3.2.8.2 Procedure
The clones were amplified by three pairs of cluster primers-C1, -C2 and -C3 (Fw and Rev) and another set of three primers: Forward-all-CP and Rev-C1, Rev-C2 and Rev-C3 (Table 3.1). The PCR products were separated on a 1% (w/v) agarose gel (Section 2.3.2). The PCR reaction mixture was prepared as per Table 3.2, and subjected to PCR cycling conditions: denaturation at 94°C for 5 min and 94°C for 45 s, annealing at 60°C for 45 s, polymerisation at 72°C for 1-2 min (25 cycles) and final extension at 72°C for 7 min.

Table 3.2 PCR reagents for amplification of CPs from cosmid library clones using cluster primers-C1, C2 and C3 (Fw and Rev), Forward-all-CP and Rev-C1, Rev-C2 and Rev-C3 primer.

PCR reagents	Volumes 25 μl	Final concentration
DNA/glycerol stocks	1.0	5-500 ng
Cluster primers-C1, -C2 and -C3 (Fw)	0.1	0.2-1 μM
Cluster primers-C1, -C2 and -C3 (Rev)	0.1	0.2-1 μM
10 × PCR buffer II	2.5	1×
25 mM MgCl$_2$	2.5	2.5 μM
10 mM dNTPs	0.5	200 μM each
Ampli*Taq*® DNA polymerase	0.125	0.625 units
dH$_2$O	18.175	

3.3   Results

3.3.1   Construction of a 2 kbp congopain gene unit library

In a previous study (Boulangé, manuscript in preparation), a *T. congolense* cosmid library had been constructed, and 15 cosmid clones had been selected by screening with CP2 coding sequence used as a probe. The existence of a *Pst*I site towards the 3' end of the three known congopain-like cysteine protease genes was exploited in the construction of a sub-cosmidic library. The 15 cosmid clones digested with *Pst*I gave bands at approximately 2 kb (single gene unit), 2.3 kb and 4 kb (likely to consist of two gene copies, one missing the *Pst*I site) (Figure 3.2). There were other bands on the gel but they did not show any given pattern.

The bands at 2 kbp were excised and cloned into pGEM®-3Zf(+). About 120 recombinant clones were obtained. All the clones were digested with *Pst*I and inserts of approximately 2 kb could be observed confirming that they were recombinants (Figure 3.3).
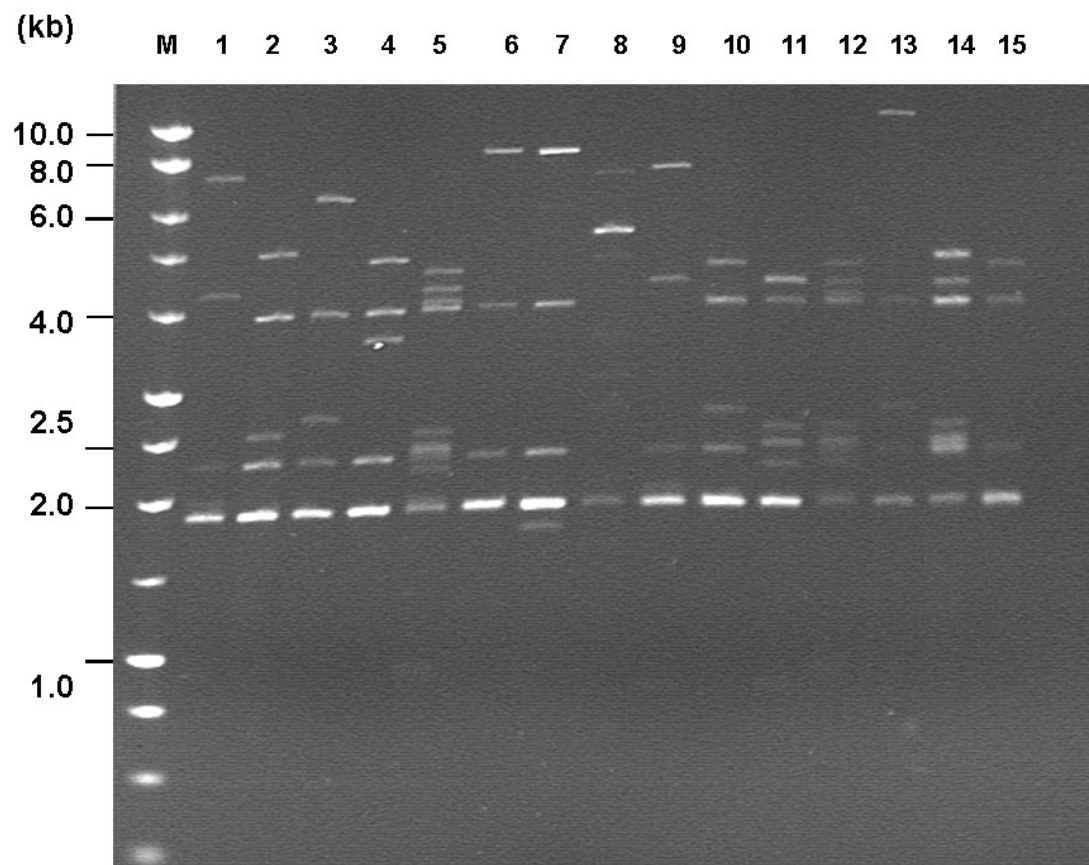
Figure 3.2 *Pst*I restriction digest of cosmid library clones 1-15. The digested samples were separated on a 0.6% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Lane M, molecular weight markers; lanes 1 – 15, cosmid library clones 1 – 15 digested with *Pst*I.
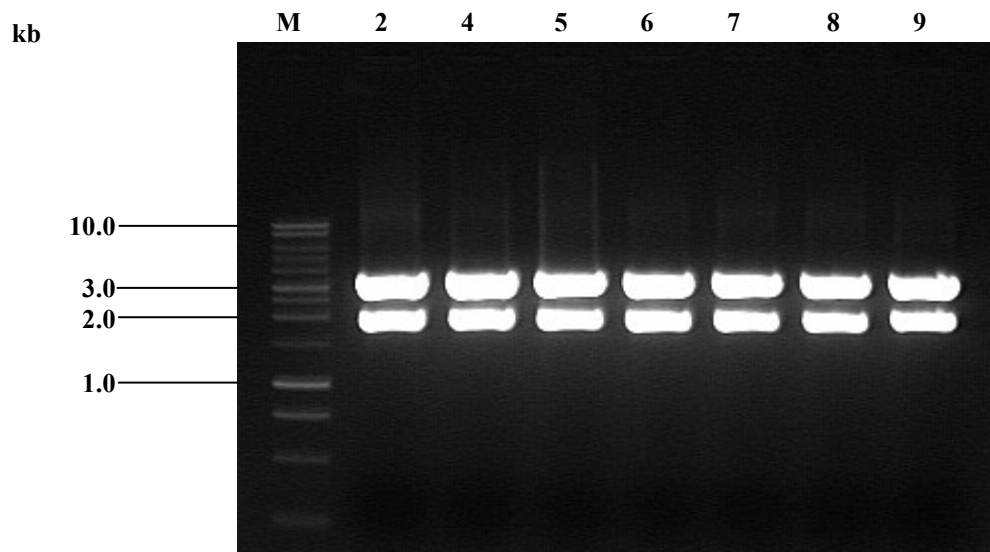
Figure 3.3 *Pst*I restriction digest of recombinant pGEM®-3Zf(+)/2 kb clones. The digested samples were separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Lane M, molecular weight markers; lanes 2, 4, 5, 6, 7, 8 and 9, corresponding 2 kb clones 2, 4, 5, 6, 7, 8 and 9 digested with *Pst*I.

3.3.2 Sequence analysis of 2 kb CP cosmid clones

Sequencing was done for 2 kb cosmid clones 1-30, using SP6 and T7 primers (Section 2.10.1) as well as internal primer Rev-new-830 (Table 3.1) to obtain the complete gene sequence. Eighteen clones were sequenced to full length, but 7 clones were only partial; the remaining 5 were either non recombinant or had more than one insert. The analysis was hence performed on 25 clones. The DNA sequences were compared with other CP genes in the database using the NCBI BLAST website (Section 2.14). All the clones matched significantly with all the three cysteine proteinase genes described in *T. congolense*; CP1 (Fish *et al*., 1995, EMBL accession number Z25813), CP2 (Jaye *et al*., 1994, EMBL accession number L25130) and CP3 (Downey and Donelson, 1999, EMBL accession no. AF139913). The DNA sequences were translated into predicted amino acid sequences using ExPASy tools and aligned by ClustalW (Figures 3.6-3.8), and showed high sequence identity (88-100%).

The signal peptides (20 amino acid long) in all 25 clones had 100% sequence identity; the propeptide was highly homologous as well with only very few amino acid substitutions. The inhibitory peptide sequence YHNGA (Lalmanach *et al*., 1998) of the propeptide (Figure 3.4, text box 1) was present and conserved in all the clones.

Since part of the catalytic domain of some of the clones was not sequenced, the variability was examined mainly around the active site cysteine residue of the enzymes (Figure 3.4). Two distinct propeptide cleavage sites were found, APEA in 22 clones and APDA in three clones (Figure 3.4, text box 2). All the clones coding for the latter proenzyme cleavage site sequence carried a serine residue at the active site instead of a cysteine (Figure 3.4, text box 4). All clones coding for APEA, on the other hand had a cysteine residue, apart from one clone coding for a serine (Figure 3.4, text box 4). An QGQ sequence at positions 19-21 (papain numbering) was present in 19 clones coding for a cysteine residue at the active site, while it was replaced by ERL in all the clones coding for a serine at the active site (Figure 3.4, text box 3). The exceptions were one clone coding for a serine residue that had a QGQ residue in place of ERL (Figure 3.4, text box 3), and only CP1 has QGA in this position. The histidine and asparagine residues forming the catalytic triad were conserved in all the clones (Figure 3.5, text boxes 1 and 2 respectively). The C-terminal extension sequences are joined to the catalytic domains by a 6-residue polyproline hinge (Figure 3.6, text box 1) and was 91-100% identical.

A phylogenetic tree based on the amino acid sequence alignment using ClustalW formed five clusters, four clones clustered together with CP2 and two more clusters of five and six clones were close to the CP2 cluster (Figure 3.7). The fourth cluster consisted of four clones with a serine residue at the active site and was very similar in sequence to CP3. The last cluster was composed of six clones closely related to CP1. Overall six clones were considered closely related to CP1, fifteen to CP2 and four clones to CP3. It should be noted however that if several clones exactly match the sequence of CP2 present in the database, none exactly match CP1 or CP3.

A certain degree of variability seems to exist within these clones mainly in the catalytic domain, with fewer differences in the amino acid sequence at the maturation site and amino acid residues forming the catalytic triad. As shown in Figure 3.8 regions represented by text boxes 1 and 2 have a higher degree of variability. The amino acid residues at these positions show the difference in the cysteine protease isoforms, clones closely related to CP1 have a different sequence compared to those closely related to both CP2 and CP3. In some instances (Figure 3.8 text box 1) clones coding for negatively charged amino acid residues (glutamate) are substituted by positively charged amino acid residues (arginine) or the amide asparagine,

while hydrophilic amino acid residues (aspartate) are replaced by hydrophobic amino acid residues (phenylalanine). Further significant differences in sequences are shown in Figure 3.8, text box 2 such as the substitution of aspartate in CP2-like clones by asparagine in CP1-like clones. The part of the catalytic domain shown in Figure 3.8 has a higher degree of variability than the part illustrated in Figure 3.5. There is very little sequence polymorphism in the C-terminal extension (Figure 3.6), unlike at the cleavage site and around the active site (Figure 3.4) of these cysteine protease clones.

```
CLUSTALW (1.83) multiple sequence alignment
                                    1                  2              3     4
Clone22     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone24     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone30     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone17     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone16     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone15     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone14     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone13     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone11     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone10     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone09     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone08     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone07     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone06     MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
CP2         MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone05     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone23     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone27     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone19     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone20     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone25     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
Clone26     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APEA VDWRKKGAVTPVKD QGQC GS C W 60
CP1         MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APPA VDWRKKGAVTPVKD QGAC GS C W 60
CP3         MSPEEFRAT YHNGA EYYAAALKRPRKVVNVSTGK APPA IDWRKKGAVTPVKD QGQC HS S W 60
Clone02     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APDA VDWRKKGAVTPVRD ERLC DS S W 60
Clone04     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APDA VDWRKKGAVTPVRD ERLC DS S W 60
Clone29     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APDA VDWRKKGAVTPVRD ERLC DS S W 60
Clone21     MSPEEFRAT YHNGA EYYAAALKRPRKVVTVSTGK APEA VDWRKKGAVTPVKD ERLC DS S W 60
            ********* *.*** *.************.***** ** * :.**********:* : * * *
```

Figure 3.4 ClustalW alignment of predicted amino acid sequences showing the N-terminus of mature CP and part of the catalytic domain of CP obtained from cosmid library clones. The clones were sequenced and the DNA sequences translated into amino acid sequence using ExPASy software and aligned by ClustalW. CP1, CP2 and CP3 are predicted amino acid sequences of cysteine protease isoforms CP1, CP2 and CP3 respectively. Text box 1, inhibitory peptide YHNGA of the propeptide; text box 2, the cleavage site between the propeptide and the central domain; text box 3, amino acid residues QGQ/QGA/ERL (positions 18-21 papain numbering) and text box 4, either cysteine or serine residues at the active site.

```
CLUSTALW (1.83) multiple sequence alignment
                               1                    2
Clone24      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
CP2          SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone08      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone11      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone20      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone07      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone17      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone10      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone22      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone06      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone16      SEHLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone13      SEHLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone01      SEHLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone14      SEHLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone30      SEHLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
CP3          SEHLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG 60
Clone05      SKQLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
Clone23      SKQLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
Clone26      SKQLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
Clone15      SKQLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
Clone19      SKQLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
Clone27      SKQLD H GVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
CP1          SKGLD H DVLLVGYDDTSKPPYWIIK N SWSKGWGEEGYIRIEKGTNQCLMKNYARSAVVSG 60
Clone25      SKQLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
Clone29      SEQLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
Clone02      SEQLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
Clone04      SEQLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
Clone21      SEQLD H GVLLVGYDDTSKPPYWIIK N SWGKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG 60
             *:  ** *  . :******************  *  **.:.********************  .  ******
```

Figure 3.5 ClustalW alignment of predicted amino sequences showing part of the catalytic domain that includes the Histidine and Asparagine of the catalytic triad of CP obtained from cosmid library clones. The clones were sequenced and the DNA sequences translated into amino acid sequence using ExPASy software and aligned by ClustalW. CP1, CP2 and CP3 are predicted amino acid sequences of cysteine protease isoforms CP1, CP2 and CP3 respectively. Text box 1, conserved histidine and Text box 2, conserved asparagine residues of the catalytic triad.

```
CLUSTALW (1.83) multiple sequence alignment

                    1
Clone5      PVPPPPPPAS TTTQEFCEGAQCQSGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 60
Clone26     PVPPPPPPAS TTTQEFCEGAQCQSGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 60
Clone19     PVPPPPPPAS TTTQEFCEGAQCQSGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 60
Clone27     PVPPPPPPAS TTTQEFCEGAQCQSGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 60
Clone23     PVPPPPPPAS TTTQEFCEGAECQSGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 60
Clone21     PVPPPPPPAS TTTQEFCEGAEC-SGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 59
Clone25     PVPPPPPPAS TTTQEFCEGAECQSGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 60
Clone29     PVPPPPPPPP TTTQEFCEGAQC-SGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 59
Clone04     PVPPPPPPPP TTTQEFCEGAQC-SGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 59
Clone02     PVPPPPPPPP TTTQEFCEGAQC-SGCTKATFPTGKCVQLSGAGSVIASCGPNNLTQIVYP 59
Clone20     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
CP2         --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone24     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone08     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone07     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone10     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone22     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone06     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone16     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone17     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone11     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone13     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSAIASCGSNNLTQIVYP 58
Clone30     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSAIASCGSNNLTQIVYP 58
Clone14     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
CP3         --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone01     --PPPPPPTP TTTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone15     PVPPPPPPAS TTTQEFCEGAECQSGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 60
CP1         PPPPPPPPAS TTTQEFCEGAECQSGCTKATFPTGKCVQFGGAGSVIASCGSNNLTQIVYP 60
                ******..  *.***:****:*  * *************:.****.*****.*********
```

Figure 3.6 ClustalW alignment of predicted amino sequences showing the C-terminal extension of CP obtained from cosmid library clones. The clones were sequenced and the DNA sequences translated into predicted amino acid sequence using ExPASy software and aligned by ClustalW. CP1, CP2 and CP3 are predicted amino acid sequences of cysteine protease isoforms CP1, CP2 and CP3 respectively. Text box 1, the polyproline hinge linking the catalytic domain to the C-terminal extension.

Figure 3.7 Phylogenetic tree constructed by ClustalW on congopain-like cysteine protease genes (clones) obtained from the CP cosmid library. CP1, CP2 and CP3 are cysteine protease isoforms CP1, CP2 and CP3 respectively. Clone numbers correspond to clones obtained from the cosmid library.

```
CLUSTALW (1.83) multiple sequence alignment

                    1                                              2
Clone27   CDTEDDGCGGG LMDNAFQWIVSSNKGNVFTEQSYPYA SKGGKMPPCNMSGK VVGANIRDH 185  ⎫
Clone19   CDTEDDGCGGG LMDNAFQWIVSSNKGNVFTEQSYPYA SKGGKMPPCNMSGK VVGANIRDH 240  ⎪  CP1
Clone26   CDTEDDGCGGG LMDNAFQWIVSSNKGNVFTEQSYPYA SKGGKMPPCNMSGK VVGANIRDH 240  ⎬  family
Clone05   CDTEDDGCGGG LMDNAFQWIVSSNKGNVFTEQSYPYA SKGGKMPPCNMSGK VVGANIRDH 240  ⎪
Clone25   CDTEDDGCGGG LMDNAFQWIVSSNKGNVFTEESYPYA SKGGKMPPCNMSGK VVGANIRDH 240  ⎪
Clone23   CDTEDDGCGGG LMDNAFQWIVSSNKGNVFTEQSYPYA SKGGKMPPCNMSGK VVGAKIRDH 240  ⎪
CP1       CDTTDYGCRGG LMDKSLQWIVSSNKGNVFTAQSYPYA SGGGKMPPCNKSGK VVGAKISGH 240  ⎭
Clone15   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPACDKSGK VVGAKIRDH 203  ⎫
Clone09   CQTNDFGCEGG LMDDAFKWIVSSNKDNVFTEQSY--- ------------ -------- 127  ⎪
Clone07   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪
Clone08   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪
CP2       CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪
Clone22   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪
Clone06   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪
Clone17   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎬  CP2
Clone11   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪  family
Clone16   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪
Clone10   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪
Clone24   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪
Clone20   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎪
Clone13   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPACDKSGK VVGAKIRDH 240  ⎪
Clone30   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPACDKSGK VVGAKIRDH 240  ⎪
Clone14   CDTNDFGCEGG LMDDAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDH 240  ⎭
CP3       CDTNDFGCGGG FSDPAFKWIVSSNKGNVFTEQSYPYA SGGGNVPTCDKSGK VVGAKIRDR 240  ⎫
Clone02   CDTREDGCGGG LMDRAFQWIVSSNKGNVFTEQSYPYA S---------- -------- 187  ⎪  CP3
Clone04   CDTREDGCGGG LMDRAFQWIVSSNKGNVFTEQSYPYA S---------- -------- 148  ⎬  family
Clone29   CDTREDGCGGG LMDRAFQWIVSSNKGNVFTEQSYPYA S---------- -------- 166  ⎪
Clone21   CDTREDGCGGG LMDRAFKWIVSSNKGNVFTEQSYPYA S---------- -------- 162  ⎭
          *:* : ** ** :* :::*******.**** :**       :          ::   :
```

Figure 3.8 ClustalW alignment of predicted amino acid sequences showing a variable region of congopain-like gene (clones) obtained from CP cosmid library

3.3.3 Screening of 2 kb CP gene unit clones by PCR

Cluster primers-C1, -C2 and -C3 (Fw) were used for screening 2 kb CP clones that were not sequenced. The three pairs of primers gave a band at approximately 250 bp, as shown in Figure 3.9, clones that gave a strong signal with either C1 or C2 or C3 were considered closely related to CP1, CP2 and CP3 respectively. Forward-all-CP and Rev-C1, Rev-C2 and Rev-C3 primers formed bands of different sizes (Figure 3.10). As shown with the cluster primers clones that gave a strong signal with the latter primers were designated as either CP1 or CP2 or CP3. Some uncertainty obtained with a set of primers could generally be lifted with the other set, as illustrated by clone 5, uncertain with Cluster-C1/2/3 set, but definite with Fw-all-CP set. Overall, both sets of primers identified half the clones as closely related to CP2, a third related to CP1 and the rest were considered close to CP3. These ratios were very similar to that observed when comparing the 25 clones that were sequenced (Section 3.3.2).
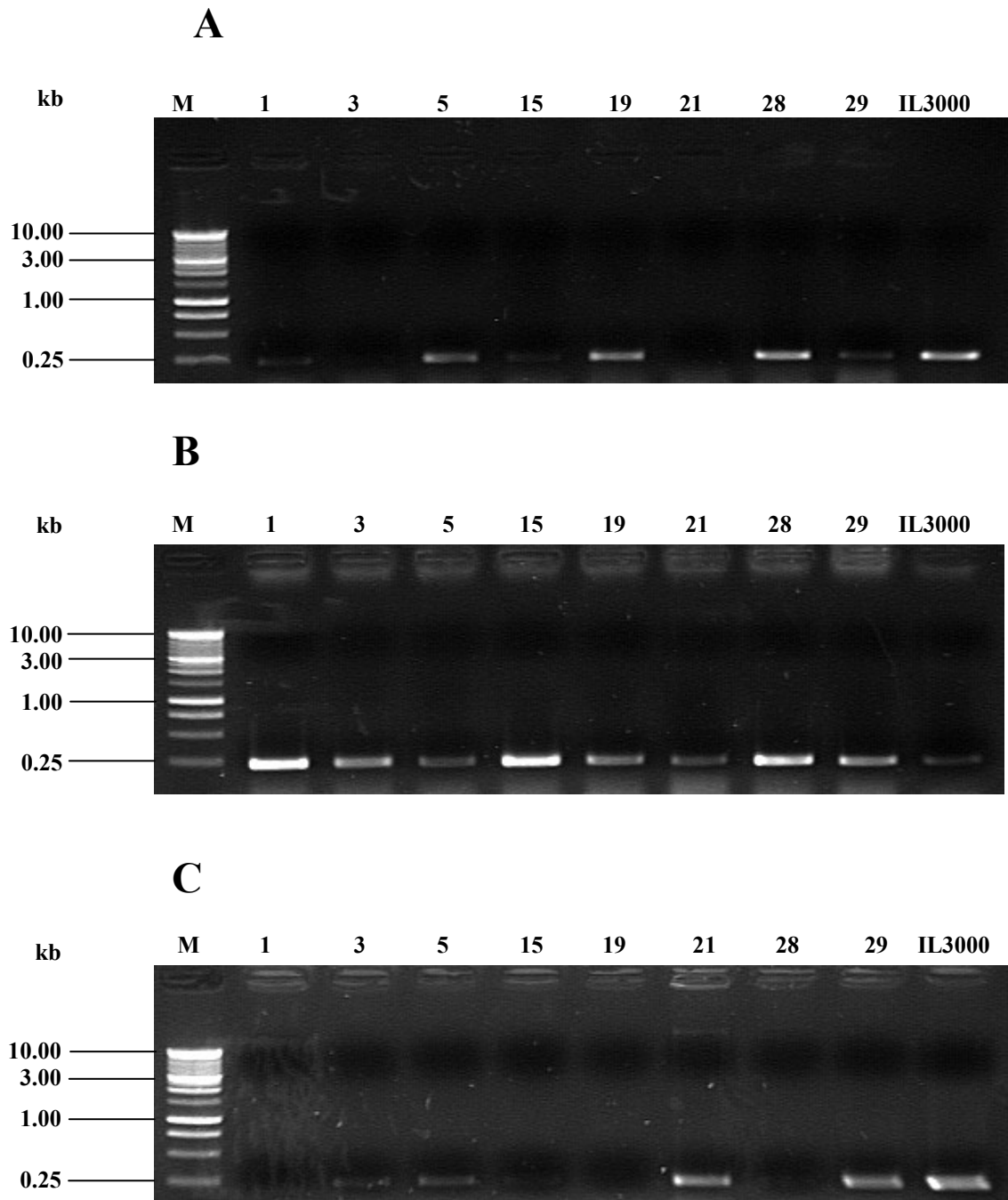
Figure 3.9 PCR on 2 kb CP cosmid library clones using cluster pairs of primers C1, C2 and C3. The digested samples were separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Lanes M, molecular weight markers; lanes 1, 3, 5, 15, 19, 21, 28, 29, clones 1, 3, 5, 15, 19, 21, 28, 29 amplified by cluster primers-C1, (Panel A), -C2 (Panel B) and -C3 (Panel C). Lanes IL 3000, *T. congolense* IL 3000 genomic DNA (positive control) amplified by cluster pairs of primers C1, C2 and C3 Fw and Rev in Panels A, B and C respectively.
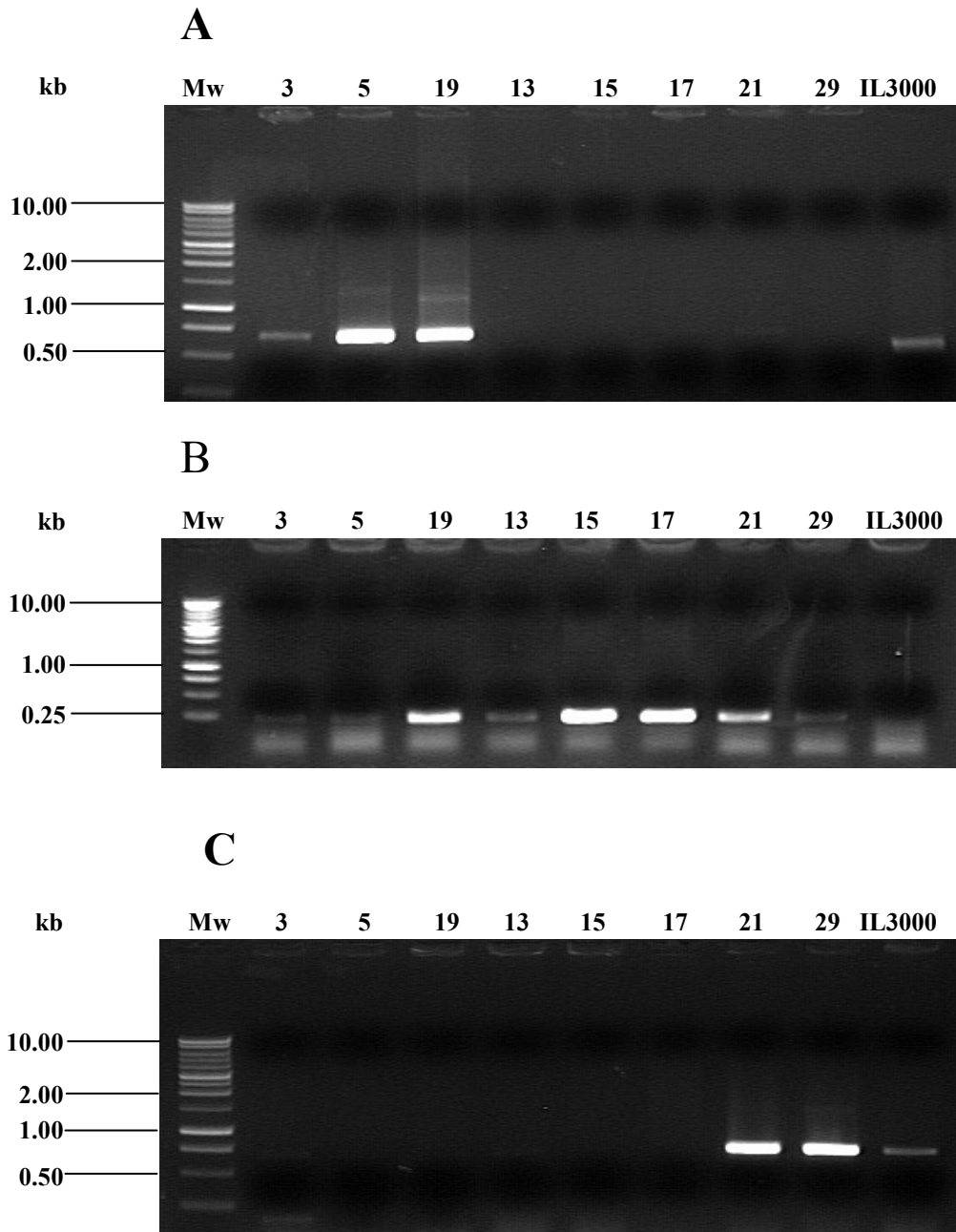
Figure 3.10 PCR on 2 kb CP cosmid library clones using Forward-all-CP and Rev-C1, Rev-C2 and Rev-C3 pairs of primers. The digested samples were separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Lanes M, molecular weight markers; lanes 3, 5, 19, 13, 15, 17, 21, 29, clones 3, 5, 19, 13, 15, 17, 21, 29 amplified by Forward-all-CP and Rev-C1 primers (Panel A), Forward-all-CP and Rev-C2 primers (Panel B) and Forward-all-CP and Rev-C3 primers (Panel C). Lanes IL 3000, *T. congolense* IL 3000 genomic DNA (positive control) amplified by C1, C2 and C3 Fw and C1 Rev, C2 Rev and C3 Rev pairs of primers in Panel A, B and C respectively.

3.4 Discussion

Approximately 120 clones of congopain-like cysteine protease were obtained after cloning the 2 kb band obtained by *Pst*I digestion of the cosmid library clones. Sequencing was done for 25 clones while the rest were screened by PCR to establish the proportions of CP1- CP2- or CP3-like cysteine protease genes. Out of the 25 sequenced clones, fifteen clustered with CP2, six clones with CP1 and four clones that code for a serine residue at the active site clustered with CP3. This was in agreement with results obtained by PCR screening using cluster primers-C1, -C2 and -C3 (Fw and Rev) and Forward-all-CP and Rev-C1, Rev-C2 and Rev-C3 primers. PCR screening identified approximately half of the clones as being close to CP2, a third as CP1 and the remaining being close to CP3. The sequenced clones were also subjected to PCR using cluster primers-C1, -C2 and -C3 (Fw and Rev), and Forward-all-CP and Rev-C1, Rev-C2 and Rev-C3 primers, the results obtained agreed with the clustering based on the phylogenic tree.

A significant degree of variability seems to exist within these clones mainly in the catalytic domain, with notable differences in the amino acid sequence at the maturation site and around the amino acid residues forming the catalytic triad. The amino acid residues at these positions show the difference in the cysteine protease isoforms. In some instances clones coding for negatively charged amino acid residues (glutamate) are substituted by positively charged amino acid residues (arginine) or the amide asparagine, while hydrophilic amino acid residues (aspartate) are replaced by hydrophobic amino acid residues (phenylalanine). Further significant differences in sequences are the substitution of aspartate in CP2-like clones by asparagine in CP1-like clones. This difference affecting the vicinity of the active site, or the overall pI of the enzyme, may affect the enzyme's activity and properties, as shown for CP1 and CP2 (Boulangé *et al*., 2001). Half of the amino acid residues substitution between C1 (CP1) and C2 (CP2), led to the former having a negative charge unlike CP2 and consequently both enzymes had different pI values and elution profiles on ion-exchange chromatography. Moreover CP2 exhibited activity over a wide range of pH unlike CP1 which was active only at acidic conditions but inactivated at neutral or alkaline pH. Lima *et al.* (2001) reported that two cysteine protease isoforms (cruzain and cruzipain 2) from *T. cruzi* with very few amino acid substitutions in the catalytic domain led to enzymes with different pH stability, substrate sensitivity and sensitivity to inhibition by inhibitors of cysteine proteases.

Of particular concern was that sequence exactly matching the CP1 or CP3 of the database where not found among these clones, while they were selected from a cDNA library. This can mean that either they are specific of the *T. congolense* strains used for their cloning, a possibility addressed in Chapter 4, or they are located outside of the cloned locus, addressed in Chapter 5.

Although there is a degree of variability in the congopain-like cysteine protease (clones), predicted amino acid sequences are 88-100% identical. Therefore it is likely that congopain based vaccine will inhibit all congopain-like cysteine protease. Due to the degree of variability among congopain-like cysteine protease clones from a single *T. congolense* IL 3000 strain, it was necessary to establish and compare their diversity in other strains of *T. congolense* as well as in other trypanosomes species as described in the next Chapter.

CHAPTER 4

VARIABILITY OF CONGOPAIN-LIKE CYSTEINE PROTEASE GENES IN TRYPANOSOMES

4.1   Introduction

Analysis of a congopain-like cysteine protease gene unit library was undertaken to study the variability of CP genes (Chapter 3). It was shown that the signal peptide and propeptide were highly conserved with very few amino acid differences. Furthermore it was shown that the catalytic domain showed variability, especially at and around the active site of the enzyme, while the C-terminal domain was more conserved. The highest degree of variability was seen in the sequences corresponding to residues cysteine to serine at the active site of the enzyme. Although the different clones obtained in the present study of *T. congolense* IL3000 strain showed at least 88% amino acid sequence identity, it was deemed necessary to the study the diversity of congopain-like cysteine protease genes in different strains of *T. congolense* to determine the extent of diversity across *T. congolense* strains. Secondly, an attempt was made to determine how different other congopain-like cysteine protease genes in *T. b. brucei* and *T. vivax* are from congopain, i.e. the interspecies variability.

To determine the variability of congopain-like cysteine protease genes in three major trypanosome species (*T. congolense*, *T. vivax* and *T. brucei*), PCR, cloning and sequencing methodology was used. Six degenerate primers were designed based on congopain-like cysteine protease genes of *T. congolense, T. vivax and T. b. brucei.* (Figure 4.1). Three degenerate species-specific primer pairs were designed based on the nucleotide sequences of the cloned congopain-like cysteine protease gene fragments. The primers were used to amplify part of the catalytic domain, but also the intergenic region. The intergenic region though not translated into protein, and hence not necessary for vaccine design, was investigated to gain insight into diversity and polymorphism.

All the PCR products were cloned into pGEM®-T Easy cloning vector, transformed into competent *E. coli* JM 109 cells and recombinant plasmid purified and sequenced. The nucleotide sequences were analysed by NCBI BLAST (www.ncbi.nlm.nih.gov/BLAST),

translated into amino acid sequences by ExPASy proteomics tools (www.expasy.org/tools) and aligned by ClustalW (www.ebi.ac.uk/ClustalW).

4.4.1 Design of degenerate and species-specific primers

At the time of designing this study, the *T. congolense* genome had not been unveiled, the only available information on the cysteine proteases of *T. congolense* was CP1 [Fish *et al.*, 1995, EMBL accession number Z25813), CP2 (Jaye *et al.*, 1994, EMBL accession number L25130 Boulangé *et al.*, 2001] and CP3 (Downey and Donelson, 1999 EMBL accession no. AF139913). This necessitated the use of the cysteine protease present in the database for *T. congolense, T. b. brucei,* and *T. cruzi* to design degenerate primers, targeting the catalytic domain and intergenic region of cysteine protease genes.

Six degenerate primers (Figure 4.1) were designed to amplify cysteine proteases in trypanosome species (*T. congolense, T. vivax and T. b. brucei*), based on an alignment of cysteine protease genes from *T. congolense* (CP1 and CP2), *T. b. brucei* and *T. cruzi* (Appendix B). Two pairs of primers CP2/CP3 and CP2/CP5 targeted the gene coding for the catalytic domain, CP2/CP3 primer pair (Table 4.1) was designed to amplify part of the propeptide and part of the catalytic domain around the active site cysteine of the enzyme. CP2/CP5 primer pair (Table 4.1) was designed to amplify part of the propeptide and almost the entire catalytic domain. The CP1/CP4 primer pair (Table 4.1) was designed to amplify a portion of the catalytic domain, the entire C-terminal extension, the intergenic region, the signal peptide and part of the propeptide (Figure 4.1). The CP1/CP6 primer pair (Table 4.1) targeted part of the C-terminal extension, the intergenic region, the signal peptide and part of the propeptide region.

After obtaining a few clones with the degenerate primers, the clones were sequenced and aligned to design species-specific primers. Clones obtained from *T. congolense* strains genomic DNA using CP2/CP3 primers were aligned (Appendix C) and new primers Tc/Tv CP2/CP3 designed. Tc CP1/CP4 primer pair was designed from alignment of DNA sequences of clones obtained using the degenerate CP1/CP4 primer pair (Appendix D). Lastly DNA sequences obtained from *T. vivax* clones amplified by the CP1/CP6 primer pair were aligned to design species-specific primers Tv CP1/CP6 (Table 4.1).
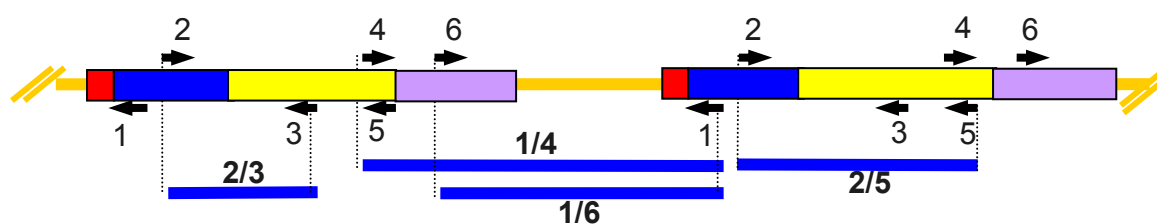
Figure 4.1 Location of degenerate primers on cysteine protease genes. Numbering 1-6, positions of primers CP1, CP2, CP3, CP4, CP5 and CP6 respectively. 2/3 part of the coding region amplified by CP2/CP3 primer pair, 2/5 part of the coding region amplified by primers CP2/CP5. 1/4 and 1/6 coding and intergenic regions amplified by CP1/CP4 and CP1/CP6 respectively. In Red: pre-peptide; in blue: propeptide; in yellow: catalytic domain; in purple: C-terminal domain; in orange: intergenic region.

4.2   Materials

87 samples of genomic DNA from *T. congolense, T. vivax, T. b. brucei, T. simiae*, *T. equiperdum* and *T. lewisi* (Table 2.3)

PCR reagents (Table 2.4)

Primers (Tables 4.1)

Table 4.1 Primers used for the amplification of *Trypanosoma* spp CPs. Degenerate primers, Y=C/T, R=A/G, M=A/C, S=C/G, H=A/C/T, V=A/C/G and N=A/C/G/T.

Name	Sequence (5'-3')	T_m (°C)
CP1	TGCYRTRCTTYTKCTTGAA	53
CP2	GCYRMGGAGGARGCRTTCCG	67
CP3	GACACKAGCATYTGCTCCGA	62
CP4	TGGGGCGAGGANGGYTACATCCGCAT	70
CP5	ATGCGGATGTARCCNTCCTCGCCCCA	69
CP6	AGYASSHVCTGCAGCGG	55
Tc CP1	GTCATCGCCTCGTGCGGCTCC	66
Tc CP4	CAAAAATAAAAAGACAAGGCCAACGGG	63
Tv CP1	AACTAT TATGCACATGAGTCCC	48
Tv CP6	CTGCAGCGGTGTACGGCTTGAGC	65
Tc/Tv CP2	CATGGAGGGTGCAAAGGAGGAGG	63
Tc/Tv CP3	GACACTAGCATYTGCTCCGA	61

4.3   Methods

Cysteine protease gene fragments were amplified by PCR as per Section 2.4.2. The PCR products were excised from the gels and purified using the Wizard® SV Gel and PCR clean-up system (Table 2.1) as described by the manufacturer, cloned into pGEM®-T Easy cloning

vector (Section 2.7.2) and transformed into competent *E. coli* JM 109 cells as per the manufacturer's instructions. The recombinant clones were screened by colony PCR using SP6 and T7 primers (Section 2.10.2) and plasmids carrying the inserts purified using Wizard® *Plus* SV Minipreps DNA purification according to the protocol supplied and sequenced at ILRI, Nairobi, Kenya.

4.4   Results

4.4.1   Amplification of cysteine protease-like genes from *T. congolense* IL 3000

PCR reactions were optimised using *T. congolense* IL 3000 genomic DNA before using other trypanosome species and strains (Figure 4.2). Primers CP1/CP4 amplified a fragment of approximately 1200 bp in size and an unexpected band could also be seen at approximately 600 bp. The second primer pair (CP1/CP6) amplified the intergenic region, resulting in a PCR product of approximately 1 kb. Primers CP2/CP3 amplified a ~ 400 bp fragment, while CP2/CP5 amplified an ~800 bp long fragment (Figure 4.2).
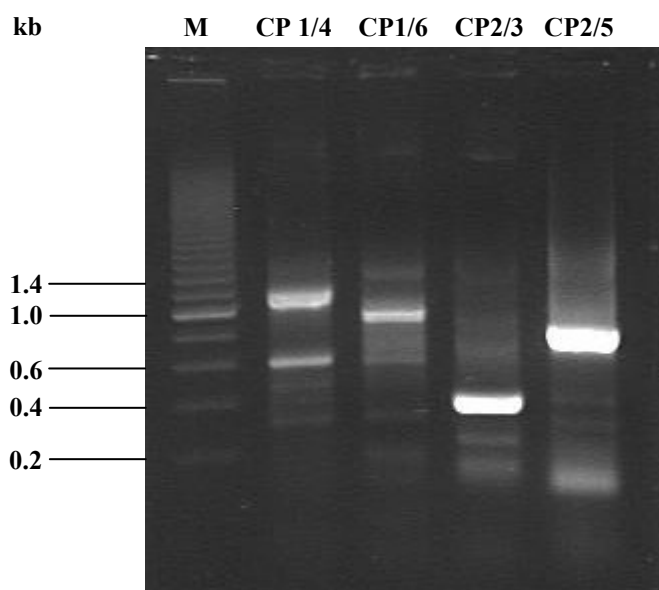
Figure 4.2 Amplification of cysteine protease genes from *T. congolense* IL3000 genomic DNA using degenerate primers. Lane M, molecular weight markers; lane CP1/4, *T. congolense* IL 3000 genomic DNA amplified by primers CP1/CP4; lane CP1/6, *T. congolense* IL3000 genomic DNA amplified by CP1/CP6 primers; lane CP2/3, PCR product from *T. congolense* IL3000 genomic DNA amplified by CP2/CP3 primers; lane CP2/5, *T. congolense* IL3000 genomic DNA amplified by CP2/CP5 primers.

4.4.2 Amplification of cysteine protease genes in other trypanosomal species

The four pairs of primers were used with genomic DNA from *T. simiae, T. vivax, T. b. brucei* and *T. congolense* as shown in Figure 4.3. The CP1/CP4 primer pair (panel A) amplified a fragment of approximately 1 kb in *T. vivax* and *T. b. brucei* (lanes 2 and 3) which was smaller than that in *T. congolense* IL 3000 (lane 4). Unlike in *T. congolense* the 600 bp was noticeably missing from *T. vivax* and *T. b. brucei* DNA. This primer pair showed no amplification of CPs from *T. simiae*. The primer set CP1/CP6 (panel B) amplified CP genes of *T. congolense*, *T. b. brucei* and *T. vivax* and the resulting PCR product was less intense compared to that obtained with CP1/CP4, similarly this primer set showed no amplification of *T. simiae* CPs. As observed earlier the PCR product obtained with this primer pair was smaller in size (approx. 800 bp) in *T. b. brucei* and *T. vivax* as compared to *T. congolense* (~1 kb), suggesting that the intergenic region of *T. congolense* is longer than that of *T. b. brucei* and *T. vivax*. Genomic DNA from all the species formed a strong band (~380 bp) upon amplification using CP2/CP3 primer pair (panel C). CP2/CP5 primer pair (panel D) amplified part of the CP genes forming a PCR product of ~ 800 bp with genomic DNA from all four trypanosome species.
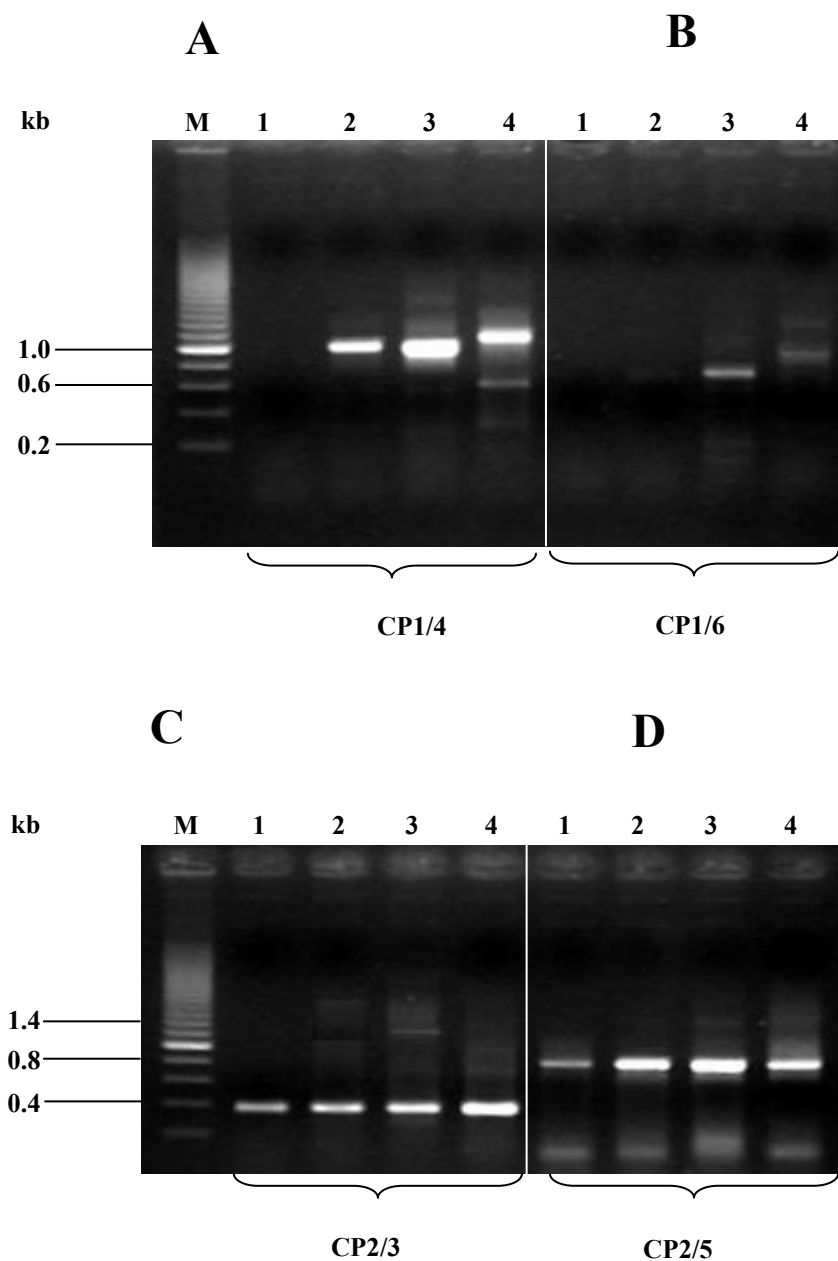
Figure 4.3 Amplification of cysteine protease genes in *T. simiae, T. vivax, T. b. brucei* and *T. congolense* using degenerate primers. The PCR products were separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Lanes 1, 2, 3 and 4, *T. simiae, T. vivax, T. b. brucei* and *T. congolense* IL3000 genomic DNA respectively and lane M, molecular weight markers. Panels A-D, amplification with primer pairs CP1/4, CP1/6, CP2/3 and CP2/5 respectively.

4.4.3 Amplification of cysteine protease genes in trypanosomes using species-specific primers

Nucleotide sequence of the gene coding for the congopain catalytic domain and the intergenic region sequence of the cloned fragments obtained using degenerate primers were aligned to design species-specific primers. The nucleotide sequences of *T. congolense* were

aligned by Sequencer 4.5 software (Gene Codes Corporation) to design species-specific primers for the catalytic domain. The new primers Tc/Tv CP2 and Tc/Tv CP3 (Table 4.1) were used to amplify and clone the gene coding for the catalytic domain of CPs from other strains of *T. congolense* and *T. vivax*. The nucleotide sequences of the intergenic region of cysteine protease of *T. congolense* strains were also aligned by Sequencer 4.5 software and new *T. congolense* specific primers were designed (Appendix C and D). Primers Tc CP1 and Tc CP4 (Table 4.1) amplified a shorter fragment of approximately 500 bp encompassing part of the C-terminal domain of CP and part of the intergenic region (Figure 4.4).
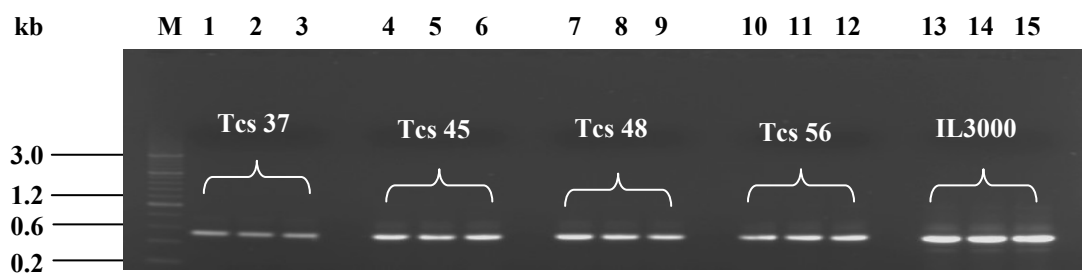
Figure 4.4 PCR on genomic DNA from *T. congolense* strains using Tc CP1 and Tc CP4. The PCR products were separated on a 1% (w/v) agarose gel and stained with ethidium bromide. (Section 2.3.2). Each PCR product was loaded into three wells and one well was skipped before loading the next PCR product. Lane 1, molecular weight markers; lanes 1-3, Tcs37; lanes 4-6, Tcs45; lanes 7-9, Tcs48; lanes 10-12, Tcs56 and lanes 13 – 15, *T. congolense* IL 3000, (see Table 2.3 for full names of *T. congolense* strains).

Nucleotide sequences of cloned intergenic regions of CPs in *T. vivax* were also aligned using the Sequencer 4.5 programme to design primers Tv CP1 and Tv CP6 (Table 4.1) and used to amplify a 800 bp fragment for all *T. vivax* strains (Tv08, Tv10, Tv13, Tv14, Tv17, Tv18, Tv21, Tv25, Tv26, Tv29, Tv30 and Tv33) as shown in Figures 4.5 A and B. Another band at approximately 400 bp was observed when *T. vivax* strains (Tv16, Tv23, Tv31) were amplified by Tv CP1 and Tv CP6 intergenic region primers (Figures 4.5 A and B) suggesting the existence of more than one locus, of differing intergenic region size, also mispriming cannot be excluded.

A
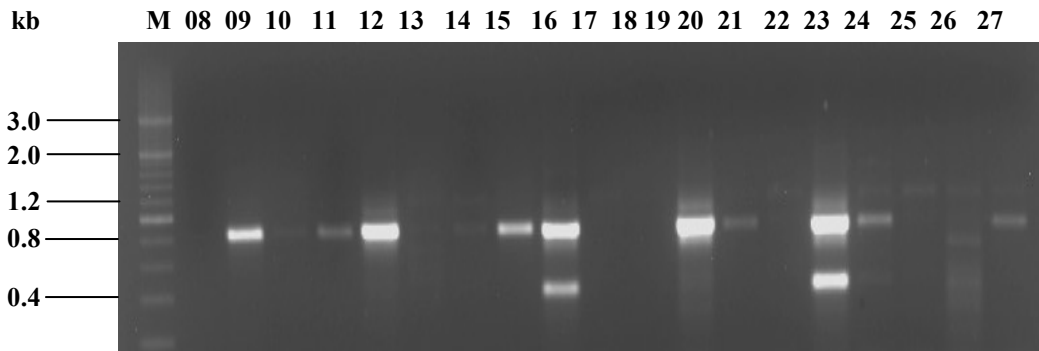
B
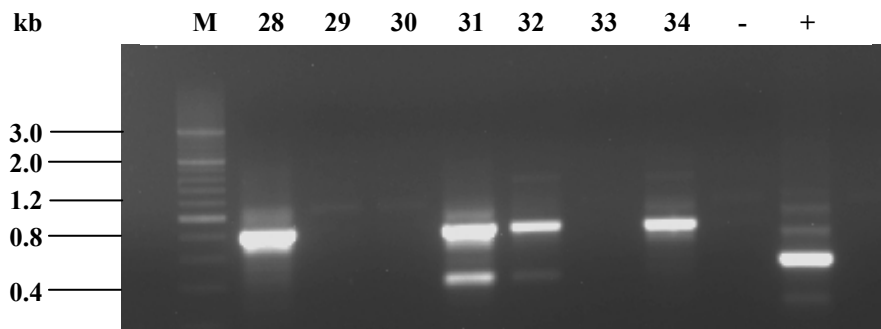
Figure 4.5 PCR on genomic DNA from *T. vivax* strains using *T. vivax* intergenic primers Tv CP1 and Tv CP6. The PCR products were separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). **(A)** Lane M, molecular weight markers; lanes 08-27, PCR on Tv08, Tv09, Tv10, Tv11, Tv12, Tv13, Tv14, Tv15, Tv16, Tv17, Tv18, Tv19, Tv20, Tv21 Tv23, Tv24, Tv25, Tv26 and Tv27 genomic DNA respectively amplified by Tv CP1 and Tv CP6 primers. **(B)** Lane M, molecular weight markers; lanes 28-34, PCR on Tv28, Tv29, Tv30, Tv31, Tv32, Tv33 and Tv34 genomic DNA respectively; lane 35, negative control (PCR master mix without DNA); lane 36, *T. congolense* IL 3000 genomic DNA (see Table 2.3 for full names of *T. vivax* strains).

4.4.4   Screening for recombinants by colony PCR using SP6 and T7 primers

After successful PCR on genomic DNA of the trypanosomes strains, the PCR products were cloned into a pGEM®-T Easy vector and white colonies screened by colony PCR using SP6 and T7 primers as shown in Figure 4.6 A and B. In the example shown on figure 4.6, four out of the 34 white colonies screened did not amplify (lane 7, 12, 27 and 34) and were considered non-recombinant, and 2 amplified a product of the wrong size (lanes 13 and 20) and were not selected (Figure 4.6 B) (for full names of *T. vivax* strains see Table 2.3). All the other colonies amplified bands of the expected sizes and were considered recombinant.

A

B
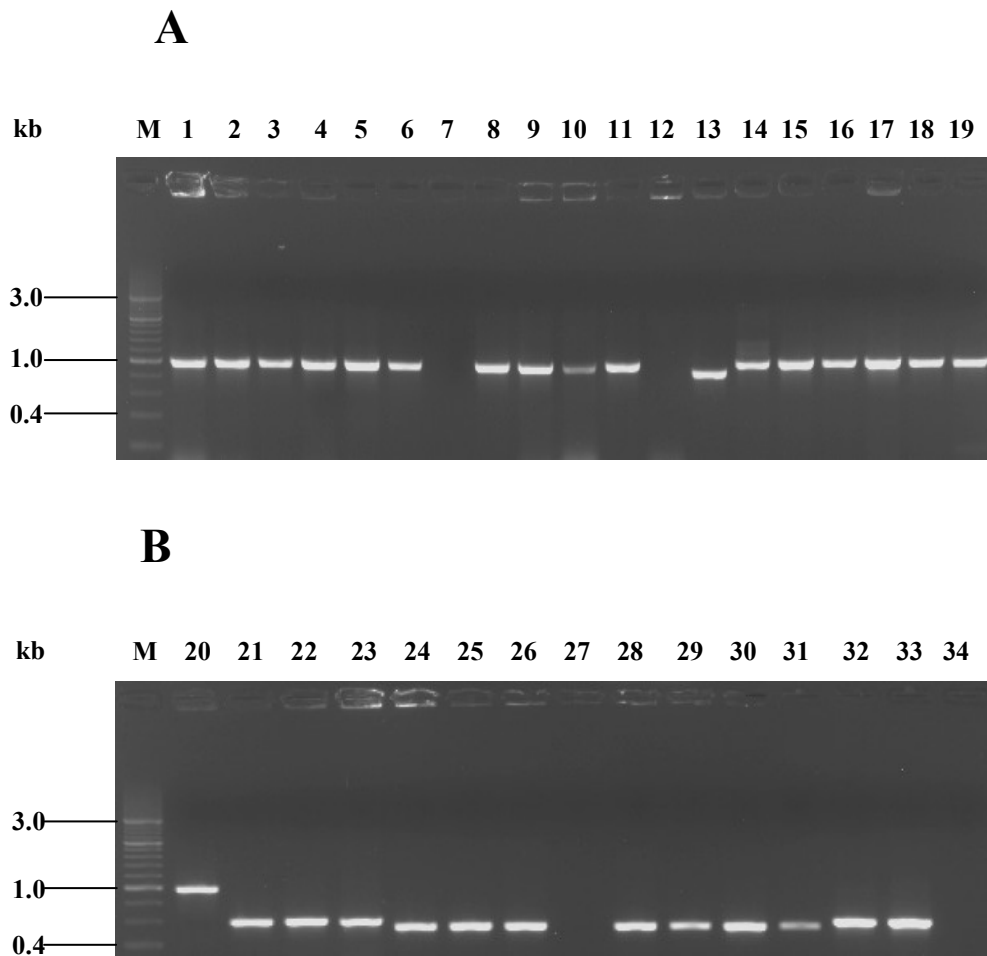
Figure 4.6 Colony PCR on *T. vivax* strains amplified by Tv CP1 and Tv CP6 and on *T. congolense* strains amplified by Tc/Tv CP2 and Tc/Tv CP3 and cloned into pGEM®-T Easy cloning vector. The PCR products were separated on a 1% (w/v) agarose gel stained with ethidium bromide (Section 2.3.2). **(A)** Lane M, molecular weight makers; colonies from lanes 1-2, Tv09; lanes 3-5, Tv12; lanes 6-8, Tv16; lanes 9-11, Tv19; lanes 12-14, Tv20; lanes 15-17, Tv28 and lanes 18-19, Tv34. **(B)** Lane M, molecular weight markers; lane 20, Tv34; lanes 21-23, 400 bp Tv16; lanes 24-26, Tcs37; lanes 27-29, Tcs45; lanes 30-31, Tcs48; lane 32, (positive control) Tcs57; lane 33, (positive control) recombinant plasmid DNA from Tcs57 and lane 34, a blue colony (negative control), (see Table 2.3 for full names of *T. vivax* and *T. congolense* strains).

4.4.5 Sequence analysis

All the nucleotide sequences were translated and aligned (Section 2.14). The cysteine protease genes amplified by both CP2/CP3 (degenerate) and Tc/Tv CP2 and Tc/Tv CP3 (species-specific) primers (Table 4.1) from *T. congolense* strain genomic DNA, showed a high degree of identity at amino acid level ranging between 79-100 %, possibly because the amplified fragment included the gene coding for part of the conserved propeptide region and only 55 amino acid residues from the more variable catalytic domain (Figure 4.7). The gene

coding for the 70 amino acid residue-long propeptide amplified by the two pairs of primers was 92-100% identical among *T. congolense* strains, while the part of the catalytic domain cloned was between 81-100% identical. The peptide sequence YHNGA in the propeptide responsible for inhibition of congopain (Lalmanach *et al*., 1998) was present in all the *T. congolense* strains (Figure 4.7, text box 1), while a similar peptide (YRNGA) was also present in all *T. b. brucei* strains cloned (Figure 4.8, text box 1).

The cleavage site between the propeptide and the catalytic domain varied with 11 out of the 18 *T. congolense* clones coding for an APEA N-terminal sequence, 4 clones coding for APPA, a single clone for APDA and 2 clones for PPMT N-terminus sequence (Figure 4.7, text box 2). The sequences APPA and APEA at the proenzyme cleavage site of cysteine protease have been reported in CP1 and CP2 respectively (Boulangé *et al*., 2001). The cysteine residue at the active site of the catalytic domain was conserved in 14 clones, 3 clones from *T. congolense* strains had a serine residue instead and included Tcf2, Tcs41 and Tcs50 (for full names of *T. congolense* strains see Table 2.3). Apart from the serine residue, one clone (Tcs65) coded for a phenylalanine residue instead of cysteine at the active site (Figure 4.7, text box 3). The only *T. simiae* strain analysed in the present study was 100% identical to *T. congolense* CP1 (Figure 4.9). The *T. simiae* clone had the peptide YHNGA in the propeptide, APPA N-terminal sequence for the mature form of the protease and a conserved cysteine residue in the catalytic triad.

The entire sequence coding for the 97 residue long C-terminal domain of the cysteine proteases in *T. congolense* was amplified by CP1/CP4, except for one strain, *T. congolense* Savannah (Tcs36) that had a 98 amino acid long C-terminal region. The C-terminal regions of these clones were 83-97% identical and are linked to the catalytic domain by six proline residues. The same clone of *T. congolense* (Tcs36) had an unusual polyproline hinge, composed of 14 proline residues linking the catalytic domain to the C-terminal region (results not shown). A complement of the 22 nucleotide polypyrimidine tract (TTTCCTTTTCTTTCTTCTTTTC) that precedes the start codon of cysteine proteases was found in all intergenic region sequences from *T. congolense* strains (Figure 4.10, text box 1).

Six clones showing sequences coding for the catalytic domain were obtained from seven strains of *T. b. brucei*, one did not amplify, and these exhibited a higher identity (97-100%) than was the case for *T. congolense* (Figure 4.8). The 71 amino acid-long propeptide was 97-

100% identical with only two amino acid differences, while the part of the catalytic domain cloned (54 amino acid long) was 98-100% identical with only one amino acid residue difference. The cleavage site (APAA) between the propeptide and the catalytic domain in *T. b. brucei* strains was conserved in all six clones (Figure 4.8, text box 2), as was the cysteine residue in the active site of the catalytic domain (Figure 4.8, text box 3). The CP amino acid sequence of a *T. equiperdum* strain was aligned with six clones of *T. b. brucei* and found to be 98-99% identical, as was the amino acid sequence from a *T. lewisi* clone (Figure 4.9). The two *T. equiperdum* and *T. lewisi* clones had the same inhibitory propeptide peptide sequence, proenzyme cleavage site and a conserved cysteine residue at the active site as was the case for *T. b. brucei* strains (results not shown).

The CP C-terminal regions in *T. b. brucei* clones were 93-100% identical and are linked to the catalytic domain by eight proline residues in four clones while clones *T. b. brucei* ILTat 1.1 and *T. equiperdum* had nine proline residues (results not shown). The intergenic region was 98-100% identical, which is higher than that in *T. congolense* (96-98%). The identity of the intergenic region in *T. vivax* was also in the same range (93-98% identity) as shown in Figure 4.13.

Cysteine protease genes from *T. vivax* were amplified using the CP2/CP3 primer pair in a few strains and gave very faint bands that necessitated to be reamplified in order to obtain sufficient material for cloning. Such fragments were cloned and sequenced. However, analysis showed that these sequences were not for *T. vivax*, but *T. congolense*. It is likely that the degenerate primers did not amplify from *T. vivax* genomic DNA, but rather picked up *T. congolense* contamination present in the laboratory. Consequently, such sequences were not included in the analysis. Time and resources being scarce, a new *T. vivax* analysis with new primers was not launched, given that the primary target of the study is *T. congolense*.

A neighbouring joining tree was generated using the predicted amino acid sequences of 27 clones (propeptide and catalytic domain, 111 amino acids long) from *T. congolense* (18), *T. b. brucei* (6), one clone from each of *T. lewisi, T. equiperdum* and *T. simiae*, two sequences from the gene database *T. vivax* and *T. cruzi*, CP1, CP2/CP3, was done. These clones were found to be 56-100% identical (not shown). Seven clusters were formed, the first consisted of two clones Tcs49 and Tcs59 strains (Figure 4.9, cluster 1), the second cluster comprised of CP1-like cysteine proteases which included the *T. simiae* clone, and two other clones from

Tcs36 and Tcs57 strains (Figure 4.9, cluster 2). The third cluster comprised CP3-like cysteine proteases and consisted of clones Tcf02, Tcs50, Tcs41 and Tcs65 (Figure 4.9, cluster 3). These clones had a serine residue instead of a cysteine at the active site apart from Tcs65 which had a unique phenylalanine residue. A fourth cluster consisted of Tcs37, Tcs45, Tcs62, Tcs66 and Tcs58 strains (Figure 4.9, cluster 4). The fifth cluster comprised of CP2-like cysteine proteases and consisted of clones from Tcs43, Tcs47, Tcs48, Tcs60, Tcs54 strains (Figure 4.9, cluster 5). The sixth cluster was comprised of clones from *T. equiperdum*, Tb05, Tb03, Tb06, Tb07, TbILtat, *T. lewisi* and TbCP23 strains (Figure 4.9, cluster 6). The seventh cluster had two sequences coming from the database, for *T. vivax* and *T. cruzi* (Figure 4.9, cluster 7) which were used as the outlaying group in this alignment. In summary, cluster one to five consisted of clones from *T. congolense* strains apart from the *T. simiae* clone which had a high identity with CP1. Whereas all the *T. congolense* clones, except Tcf02 belonged to the Savannah subgroup, it would have been interesting if several Forest and Kilifi subgroups strains of *T. congolense* were included in the present study. The sixth cluster had all the *T. brucei ssp* clones including *T. equiperdum* and *T. lewisi*.

The DNA sequences for the intergenic region of CP genes obtained from *T. brucei, T. congolense* and *T. vivax* strains showed an unexpected high conservation rate at 93-98% identity (data not shown). In comparison, the intergenic region of the cysteine proteases genes was found to be more conserved than the gene coding region, which goes against the admitted paradigm. The region was 93-98% identical in *T. vivax*, 96-98% identical in *T. congolense* and 98-100% identical for *T. b. brucei*, while the catalytic domain was 81-100% identical in *T. congolense* and 98-100% identical in *T. b. brucei*. It appear generally that cysteine proteases are more conserved in *T. b. brucei* than in *T. congolense* and *T. vivax*

```
CLUSTALW (1.83) multiple sequence alignment
                                                                      1                    2                3
Tcs37          --------------MEGAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 46
Tcs45          --------------REGAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 46
Tcs43          --------------MEGAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 46
Tcs47          --------------MEGAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 46
Tcs60          --------------MEGAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 46
Tcs48          --------------MEGAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 46
Tcs54          AEEEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 60
Tcs62          AEEEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDMPPEEFRAT YHNGA EYYAAALKR 60
Tcs49          AEEEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDVSPEEFRAT YHNGA EYYAAALKR 60
Tcs59          AAEEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDVSPEEFRAT YHNGA EYYAAALKR 60
Tcs58          AAEEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 60
Tcs66          AAEEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 60
Tcs57          -AEEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 59
Tcs36          AEEEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 60
Tcs50          --------------MEGAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 46
Tcs65          --EEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 58
Tcs41          AEEEAFRFRFFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 60
Tcf2           -EEEAFRFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKR 59
                 * ********************:.******* ***** *********

Tcs37          PRKVVNVSTGK APEA VDWRKKGAVTPVKDQGRCGS C WAFSAIGNIEGQWKIAGHELTSLS 106
Tcs45          PRKVVNVSTGK APEA VDWRKKGAVTPVKDQGRCGS C WAFSAIGNIEGQWKIAGHELTSLS 106
Tcs43          PRKVVTVSTGK APEA VDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKVAGHELTSLS 106
Tcs47          PRKVVTVSTGK APEA VDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKVAGHELTSLS 106
Tcs60          PRKVVTVSTGK APEA VDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKVAGHELTSLS 106
Tcs48          PRKVVNVSTGK APPA VDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKVTDHNLTSLS 106
Tcs54          PRKVVNVSTGK APEA VDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKVTGHNLTSLS 120
Tcs62          PRKVVNVSTGK APEA VDWRKKGAVTPVKDQGRCGS C WAFSAIGNIEGQWKIAGHELTSLS 120
Tcs49          PRKVVTVSTGK APDA VDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKIAGHELTSLS 120
Tcs59          PRKVVNVSTGK APPA IDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKIAGHEPTSLS 120
Tcs58          PRKVVNVSTGK APEA VDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKIAGYELTSLS 120
Tcs66          PRKVVNVSTGK APEA VDWRKKGAVTPVKDQGRCGS C WAFSATGNIEGQWKIAGHELTSLS 120
Tcs57          PRKVVTVSTGK APPA VDWRKKGAVTPVKDQGSCGS C WAFAATGNIEGQWKIAGHELTSLS 119
Tcs36          PRKVVNVSTGK APPA VDWRKKGAVTPVKDQGKCGS C WAFSATGNIEGQWKVAGHELTSLS 120
Tcs50          PRKVVNVSTGR PPMT VDWRKKGAVTPVKDQGKCDS S WAFSATGNIEGQWKIAGHELTSLS 106
Tcs65          PRKVVNVSTGE APPA IDWRKKGAVTPVKDEGKCDS F WAFSAIGNIEGQWKIAGHELTSLS 118
Tcs41          PRKVVNVSTGK APEA VDWRKKGAVTPVKDQGKCDS S WAFTVIGNIEGQWKIAGHELTSLS 120
Tcf2           PRKVVNVSTGR PPMT VDWRKKGAVTPVKDQGKCDS S WAFSATGNIEG------------- 106
                *****.*****.  .* :   :.************:* *.*   ***:. *****

Tcs37          EQMLVS 112
Tcs45          EQMLVS 112
Tcs43          EQMLVS 112
Tcs47          EQMLVS 112
Tcs60          EQMLVS 112
Tcs48          EQMLVS 112
Tcs54          EQMLV- 125
Tcs62          EQMLV- 125
Tcs49          EQMLV- 125
Tcs59          EQMLV- 125
Tcs58          EQMLV- 125
Tcs66          EQMLV- 125
Tcs57          EQMLV- 124
Tcs36          EQMLV- 125
Tcs50          EQMLVS 112
Tcs65          EQMLVS 124
Tcs41          EQMLV- 125
Tcf2           ------
```

Figure 4.7 ClustalW alignment of predicted amino acid sequences of the propeptide and part of catalytic domain of congopain-like cysteine protease genes in *T. congolense* amplified by CP2/CP3 primers. Tcf2, Tcs36, Tcs37, Tcs41, Tcs43, Tcs45, Tcs47, Tcs48, Tcs49, Tcs50, Tcs54, Tcs57, Tcs58, Tcs59, Tcs60, Tcs62, Tcs65 and Tcs66 (see Table 2.3 for full names of strains). Text box 1, the inhibitory peptide of the propeptide in *T. congolense*; text box 2, the proenzyme cleavage site of CPs in *T. congolense*; text box 3, cysteine or serine or phenylalanine residues at the active site.

```
CLUSTALW (1.83) multiple sequence alignment


                                                                          1
Tb6            AEEEAFRFRAFEENMEQAKIQAAANPYATFGVTPFSDMTREEFRAR YRNGA SYFAAAQKR 60
Tb7            AEEEAFRFRAFEENMEQAKIQAAANPYATFGVTPFSDMTREEFRAR YRNGA SYFAAAQKR 60
Tb3            AEEEAFRFRAFEENMEQAKIQAAANPYATFGVTPFSDMTREEFRAR YRNGA SYFAAAQKR 60
Tb5            AEEEAFRFRAFEENMEQAKIQAAANPYATFGVTPFSDMTREEFRAR YRNGA SYFAAAQKR 60
TbCP23         AKEEAFRFRAFEENMEQAKIQAAANPYATFGVTPFSDMTREEFRAR YRNGA SYFAAAQKR 60
TbILtat        AKEEAFRFRAFEENMEQAKIQAAANPYATFGVTPFSDMTREEFRAR YRNGA SYFAAAQKR 60
               *:******************************************** ***** *********

                                2                             3
Tb6            LRKTVNVTTGR APAA VDWREKGAVTPVKDQGQCGS C WAFSTIGNIEGQWQVAGNPLVSLS 120
Tb7            LRKTVNVTTGR APAA VDWREKGAVTPVKDQGQCGS C WAFSTIGNIEGQWQVAGNPLVSLS 120
Tb3            LRKTVNVTTGR APAA VDWREKGAVTPVKDQGQCGS C WAFSTIGNIEGQWQVAGNPLVSLS 120
Tb5            LRKTVNVTTGR APAA VDWREKGAVTPVKDQGQCGS C WALSTIGNIEGQWQVAGNPLVSLS 120
TbCP23         LRKTVNVTTGS APAA VDWREKGAVTPVKDQGQCGS C WAFSTIGNIEGQWQVAGNPLVSLS 120
TbILtat        LRKTVNVTTGR APAA VDWREKGAVTPVKDQGQCGS C WAFSTIGNIEGQWQVAGNPLVSLS 120
               **********  ****  ******************* *  **:*******************


Tb6            EQMLV 125
Tb7            EQMLV 125
Tb3            EQMLV 125
Tb5            EQMLV 125
TbCP23         EQMLV 125
TbILtat        EQMLV 125
               *****
```

Figure 4.8 ClustalW alignment of predicted amino acid sequences of the propeptide and part of catalytic domain of congopain-like cysteine protease genes in *T. b. brucei* amplified by CP2/CP3 primers. Tb3, Tb5, Tb6 and Tb7 (see Table 2.3 for full name of strains), TbCP23 (*T. b. brucei*) and TbILTat (*T. b. brucei* IL Tat 1.1). Text box 1, the inhibitory peptide in the propeptide in *T .b. brucei;* text box 2, the amino acid residues at the pro-region cleavage site of cysteine proteases in *T. b. brucei*; text box 3, the cysteine residue at the active site.
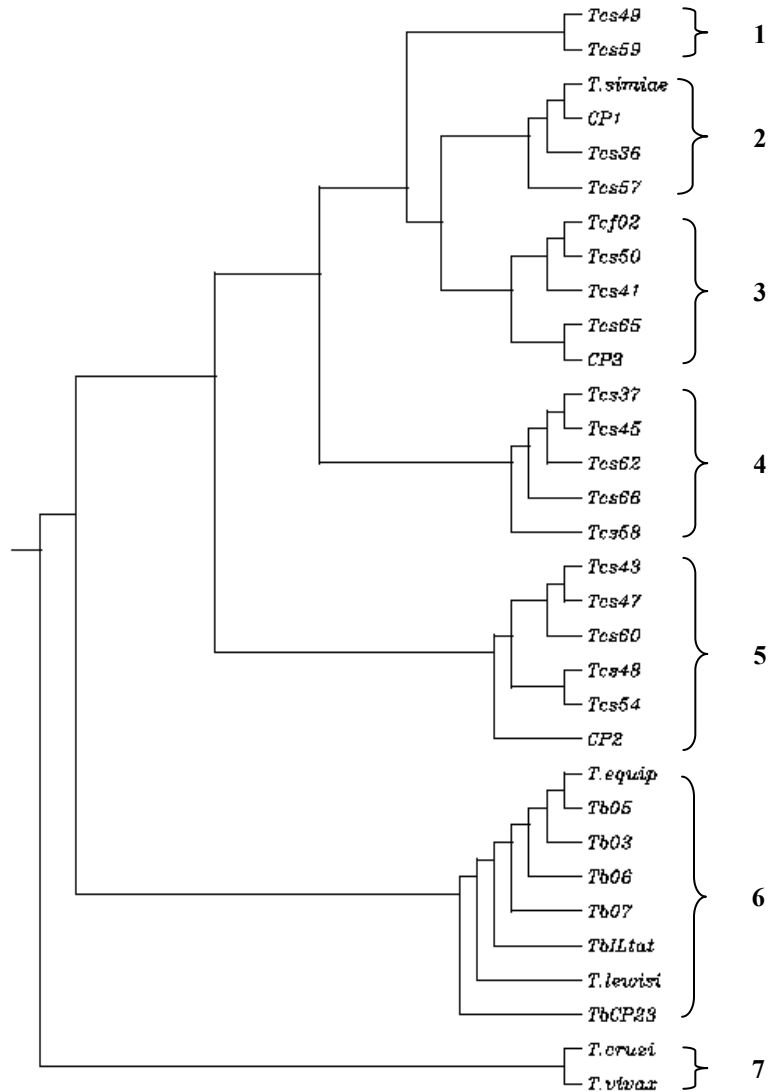
Figure 4.9 Neighbouring Joining tree generated by ClustalW alignment of 111 amino acid-long fragments of the cysteine protease genes in trypanosomes amplified by CP2/CP3 and Tc/Tv CP2 and Tc/Tv CP3 primers. Cluster 1, clones from Tcs49 and Tcs59 strains; cluster 2, clones from Tcs36 and Tcs57 strains, *T. simiae* and CP1; cluster 3, clones from Tcf02, Tcs50, Tcs41, Tcs65 strains and CP3; cluster 4, clones from Tcs37, Tcs45, Tcs62, Tcs66 and Tcs58 strains; cluster 5, clones from Tcs43, Tcs47, Tcs48, Tcs60, Tcs54 strains and CP2; cluster 7, clones from *T. vivax* and *T. cruzi*. T. equip (*T. equiperdum*), sequences obtained by amplifying and cloning genomic DNA provided by Dr. Majiwa of ILRI, Nairobi, Kenya, *T. vivax*, sequence obtained from http://www.genedb.org/genedb/tvivax/blast.jsp accession number tviv521g07.p1k_2, *T. cruzi* sequence obtained from http://www.ncbi.nlm.nih.gov/blast/Blast, accession number AAC00067.1. (for full names of strains see Table 2.3).

4.5 Discussion

Cysteine protease genes from *T. congolense, T. brucei, T. vivax, T. simiae* and *T. equiperdum* were amplified by PCR using degenerate and species-specific primers, cloned and sequenced. The PCR targeted both the gene coding and the intergenic regions of congopain-like cysteine protease genes. Out of the two pairs of primers (CP2/CP3 and CP2/CP5) for amplifying the coding region of cysteine proteases gene, the former was preferred because it amplified part of the propeptide (which is relatively conserved among *T. congolense*), and part of the catalytic domain especially around the active site cysteine residue. The primer pair (CP2/CP5) amplified a region spanning part of the propeptide till the end of the central domain, a region most interesting for our analysis. The two pairs of primers designed to amplify the intergenic region, CP1/CP4 primer pair was best suited for *T. congolense* and *T. b. brucei* strains and hence used throughout the study, whereas CP1/CP6 primer pair worked best for *T. vivax* strains. In comparison, it was more difficult to amplify CP genes from *T. vivax*, than their *T. congolense* and *T. b. brucei* counterparts, consequently there was contamination of the *T. vivax* PCR product with *T. congolense* IL 3000 genomic DNA during reamplification, and hence clones obtained from these PCR products were found to be identical to those from *T. congolense* strains. The cysteine proteases from *T. vivax* were found later to be distantly related to those from *T. congolense* and *T. b. brucei* and this probably explains why it was difficult to amplify the *T. vivax* CP genes using degenerate primers.

All the primers gave PCR products of the expected size, with the exception of the CP1/CP4 primer pair which formed an extra band at approximately 600 bp in addition to the expected 1.2 kb band. This particular band was only seen in *T. congolense* strains. Upon cloning the 600 bp PCR product, it was found not to be related to a cysteine proteases gene, and assumed to be some sequence that may share some sequence identity with the cysteine proteases in the region where these primers were designed. Two major features were shown by amplification of the intergenic region of cysteine proteases, i.e. firstly the region was found to be polymorphic in size as it was shorter in *T. b. brucei* and *T. vivax* than in *T. congolense* strains. Secondly, amplification of this region in *T. vivax* strains resulted in two bands suggesting the existences of two loci of cysteine protease genes in the *T. vivax* genome.

Species-specific primers were preferred to the original degenerate primers and were used thereafter in the present study. The switch from CP2/CP3 to Tc/Tv CP2 and Tc/Tv CP3

proofed very useful because these primers were better suited to amplify the target region. The intergenic region primers were also changed from CP1/CP4 to Tc CP1 and Tc CP4, which amplified a shorter fragment (approximately 550 bp) from *T. congolense* strains, unlike the previous 1.2 kb region. The C-terminal extension is a characteristic feature of trypanosomal cysteine proteases (Mottram *et al*., 1989; Pamer *et al*., 1990; Souza *et al*., 1992; Traub-Cseko *et al*., 1993; Stoka *et al*., 2000). The polypyrimidine tract previously described by Matthews *et al*. (1994); Fish *et al*. (1995) was found in the intergenic region just before the start codon of cysteine proteases from all *T. congolense* strains analysed. The CP1/CP6 primer pair was also replaced with *T. vivax* specific primers Tv CP1 and Tv CP6 which amplified an approximately 800 bp region encompassing the intergenic part of cysteine protease gene in *T. vivax* strains.

In the previous Chapter it was shown that the majority of the clones have a cysteine residue in the active site, likewise the majority of the clones obtained from the different strains cloned had cysteine residues. Interestingly a few clones had a serine residue instead of a cysteine, apart from one clone from Tcs65 which had a phenylalanine residue, a feature common to a group of cysteine proteases from *Plasmodium* called the Serine Repeat Antigen (SERA) which possess a central domain that show homology to the papain family of cysteine proteases, but exhibit an unusual cysteine to serine substitution at the active site (Bzik *et al*., 1988; Kiefer *et al*., 1996; Gor *et al*., 1998; Hodder *et al*., 2003). A unique proenzyme cleavage site (PPMT) between the propeptide and the catalytic domain was found in two clones Tcf02 and Tcs50, both of which possess a serine residue instead of a cysteine in the active site.

The alignment of the clones consisting of 111 amino acid-long fragment (encompassing part of the propeptide and the catalytic domain, and generated using CP2/CP3 primer pair) showed that there was a high degree of identity among clones. Only clones from *T. vivax* and *T. cruzi* were quite distantly related to the other clones, with *T. cruzi* and Tcf02 clones giving the least identity (56%). Although clone Tcf02 carried an active site serine residue, the fact that a smaller fragment (92 amino acids long) was analysed might explain why it gave the least identity with various clones (72%). All the other clones showed a high identity (74-100%); this feature is an indication that congopain is not very different from other cysteine proteases of trypanosomes and therefore an immunogen that would induce antibodies that inhibit congopain activity would most likely also act on these cysteine proteases. The

biological significance of the mutants with a replacement of their active site cysteine by a serine is puzzling, in the sense that it is far too common, and far too widespread amongst strains, to be fortuitous. As was done with SERA genes of *P. falciparum* (Miller *et al.*, 2002)*, it would be necessary to establish the expression these mutants, and also disrupt their genes through RNA interference to determine their role in host-parasite interactions. In *P. falciparum* SERA-4 and SERA-5 (with a serine residue at the active site) and SERA-6 (with a cysteine residue at the active site) were found to be highly expressed, and the fact they could not be disrupted by gene knock-out is an indication of the important role they play in maintenance of the erythrocytic cycle (Miller *et al.*, 2002). These aspects are being addressed in our laboratory but fall outside the scope of the present study.

Another interesting feature is that Neighbouring Joining trees formed upon alignment of either the gene coding region or the intergenic region sequences, placed some strains in different positions for example *T. congolense* strains Tcs57 and Tcs58 were shown to be very close when their intergenic regions are compared, but were further apart when their gene coding regions were compared. It can be suggested that since many copies of cysteine protease genes exist in the genome, it is possible that different cysteine protease genes were amplified by the two pairs of primers used, therefore it might not be ideal to compare the two regions phylogenetically. The main conclusion to draw from this chapter is that in *T. congolense*, the level of variability that we had met in one strain, IL 3000, is basically the same as one meets across *T. congolense* strains from various origins, at least from the *Savannah* group. Therefore, conclusions drawn from this IL 3000 strain as per gene organisation and variability of congopain-like CP genes would be valid for all strains.

As discussed in Chapter 3, where a 2 kb gene unit library of CP genes construction from cosmid clones was described, it has emerged that some of the cysteine protease genes are scattered outside of the main locus, and therefore restriction endonucleases which restrict within the gene omit genes outside the locus. A restriction endonuclease which restricts the gene in the intergenic region was used to construct a subgenomic library which would include the scattered copies of cysteine protease genes. This is reported in Chapter 5.

The results presented so far showed that several clones of congopain-like cysteine proteases were found to contain an unusual substitution of the cysteine residue of the catalytic triad of cysteine proteases with a serine residue. It was not clear whether the cysteine proteases

containing the serine residue are expressed and their role in pathogenesis is unknown. As mentioned above, investigating the significance of these mutations in host-parasite relationships will require reverse genetics. The study of the expression of these cysteine proteases at the mRNA level however is an easier task that was undertaken using reverse PCR technology, and is also reported in Chapter 5.

CHAPTER 5

CONSTRUCTION AND SCREENING OF A 2 KBP SUBGENOMIC LIBRARY. INSIGHT INTO THE EXPRESSION OF MUTANT GENES.

5.1   Introduction

A gene library is a collection of cloned fragments which together represent an entire genome of a particular organism (Dale and von Schantz, 2002), while a subgenomic library generally represents an entire chromosome or gene locus, or, as in the present study, a size-restricted portion of the genome. The genomic DNA is fragmented into smaller fragments that are easily cloned into a suitable vector. Fragments to be cloned are generated using restriction enzymes for complete or partial digest of genomic DNA or by exposing the genomic DNA to mechanical shearing. A DNA probe is the most common method used to identify clones harbouring the gene of interest in a library. In certain cases PCR can also be used, because it is less tedious, time saving and has a high stringency as compared to the use of probes (Israel, 1993).

The genomic organisation of congopain was studied by southern blotting using a congopain coding sequence as a probe. Congopain was found to exist at a single locus consisting of 12-14 copies of cysteine protease genes in a tandem array. It was also observed that longer exposure time of the blots resulted in additional bands to that of the congopain gene locus, leading to the hypothesis that other congopain-like genes could be scattered in the genome outside of the main locus (Boulangé manuscript in preparation). It is this hypothesis that led to the construction of a subgenomic library of 2 kb fragment using a restriction endonuclease that cuts the gene in the intergenic region so that other genes outside of the main locus of congopain would be included in the library. Restriction sites however are scarce in the intergenic region, partly because of its small size, about 500 bp, and partly because it is AT-rich. Actually, only one enzyme, *Mlu*I, was adequate, in the sense that it had a conventional recognition site, i.e. 6-cutter and palindromic, and cuts in the recognition sequence, and generates cohesive ends. Recognition sites for *Mlu*I are usually not available in the polycloning site of commonly used plasmids. A cloning vector was therefore modified by introducing a *Mlu*I site.

The catalytic triad of the papain-like cysteine proteases consists of cysteine, histidine and asparagine. (Storer and Ménard, 1994). However the existence of congopain-like cysteine protease genes having a serine residue in the place of the cysteine residue (that defines the CP3 family) reported before (Downey and Donelson, 1999, EMBL accession no. AF139913) raised concerns whether these are true genes or pseudogenes. Apart from the serine residue at the active site, CP3s also differ in their proenzyme cleavage sites resulting in N-terminal sequences varying among APEA, APPA, APDA, and PPMT. It is not clear whether the corresponding proteases are expressed in trypanosomes and what their role in host-parasite interactions might be. Boulangé *et al*. (2001) reported two cysteine protease isoforms (CP1 and CP2) from *T. congolense* with a small number of amino acid differences in their catalytic domains, yet having different enzymatic characteristics that could lead to different roles in the parasite. It was therefore of interest to determine whether these congopain-like cysteine protease genes (coding for serine instead of cysteine at the active site) are expressed in trypanosomes. Given the high level of identity of the different congopain-like gene product deduced from the cloned sequences, it would have been very difficult to study the expression of these variants at protein level. This would have necessitated highly specific anti-peptide antibodies, or the use of Mass-Spectrometry and 2-Dimensional separation. For this reason, we studied the expression at messenger level by RT-PCR on mRNA isolated from *T. congolense* IL 3000 parasites.

The RT-PCR was performed using primers Forward-all-CP and Rev-C1, Rev-C2, Rev-C3 (Table 3.1) used in Section 3.2.7. In addition, we designed Rev-C4 (Section 5.8.1), which is similar to Rev-C3, which is specific of the CP3 family, but based on the sequence of the CP3-like protease sequence present in the database (Downey and Donelson, 1999 EMBL accession no. AF139913), while Rev-C3 is based on the sequence of the CP3s cloned from the 2 kbp gene-units library generated from cosmids (Chapter 3). We used also in this study cluster primers-C1, -C2 and -C3 (Fw and Rev) (Table 3.1) used in Section 3.2.7, that generate different-sized fragment according to the targeted CP gene family.

5.2   Construction of the pMluI cloning plasmid

A mentioned above, *Mlu*I is the only adequate enzyme whose restriction site is present in the intergenic region of congopain gene. A cloning vector with an *Mlu*I restriction site was thus required. A few cloning vectors have an *Mlu*I restriction site, but they are rare, hard to

procure, and very expensive. This led to the construction of a vector containing this restriction site. Two methods of constructing the vector with an *Mlu*I restriction site were used, i.e. site-directed mutagenesis of the multiple cloning sites (MCS) of pGEM®-11Zf(+) cloning vector and cloning of annealed primers with the *Mlu*I restriction site into the pGEM®-11Zf(+) MCS.

Site-directed mutagenesis involves annealing of selection oligonucleotide and mutagenic oligonucleotide to denatured pGEM®-11Zf(+) cloning vector, followed by synthesis of the mutant strand by T4 DNA polymerase and ligation of the two oligonucleotides by T4 DNA ligase. The ligated oligonucleotides are transformed into *E. coli* strain BMH 71-18 *mut*S and cells are grown in the presence of GeneEditor™ Antibiotic selection to select clones containing the mutant plasmid. These are isolated by minipreps and transformed into competent *E. coli* JM 109 cells. Mutant plasmids can then be screened using IPTG and X-gal indicator plates or by targeting restriction sites within mutagenic primers with the appropriate endonucleases.

For the second method used, two complimentary oligonucleotides were designed, bearing a *Mlu*I site in their midst, and showing once annealed protruding ends such as those generated by *Eco*RI on one side, and *Xba*I on the other, so that the two oligonucleotides correspond to a small double-stranded DNA fragment that can be directionally cloned in a vector such as pGEM®-11Zf(+) cut with *Eco*RI and *Xba*I.

5.2.1 Alkaline denaturation of pGEM®-11Zf(+)

Alkaline denaturation refers to breaking down of the hydrogen bonds in a double stranded molecule (DNA template) to make it single stranded, by use of an alkaline solution such as NaOH. As selection oligonucleotide is an oligonucleotide provided in the Gene editor *in vitro* site-directed mutagenesis system, and encodes the mutations which alter the ampicillin resistance and hence create a new additional resistance to GeneEditor™ Antibiotic selection mix. Mutagenic primer is an oligonucleotide that incorporates the desired mutation; it is complimentary to the target template except the area of mismatch. Both primers have to hybridize at the same strand but in the opposite directions. The pGEM®-11Zf(+) vector was denatured in the presence of NaOH, to make it single stranded for hybridisation with mutagenic and selection oligonucleotides.

5.2.1.1 Materials

pGEM®-11Zf(+). (Provided in Gene editor *in vitro* site-directed mutagenesis system, Table 2.1)

10 M NaOH. NaOH pellets (4 g) were dissolved in dH$_2$O (10 ml)

0.5 mM EDTA. EDTA (1.861 g) was dissolved in dH$_2$O (8 ml), adjusted to pH 8.0 using 1 M NaOH, made up to 10 ml with dH$_2$O and autoclaved (121°C, 30 min, RT)

2 M NaOH, 2 mM EDTA. NaOH (10 M) (2 ml), 0.5 mM EDTA (40 µl) were mixed with dH$_2$O (7.96 ml).

5.2.1.2 Procedure

pGEM®-11Zf(+) (1 µg) was denatured by addition of 2 M NaOH/2 mM (2 µl) EDTA and the reaction made up to 20 µl with dH$_2$O, and the reaction carried out according to the manufacturer's instructions.

5.2.2 Hybridization of denatured pGEM®-11Zf(+) and selection of mutagenic oligonucleotides

Hybridisation involves the annealing of the selection and mutagenic oligonucleotides to the target after denaturation; it occurs at specific temperatures depending on the composition of the mutagenic primer. Generally the annealing step is done at 75°C and the reaction is allowed to cool slowly to 37°C.

5.2.2.1 Materials

Denatured pGEM®-11Zf(+)(Section 5.2.1.2)

Mutagenic primer (F3): GCCGAGCTCGAACGCGTCGACCTCGAGG, T$_m$ (70.2°C)

Selection oligonucleotides (bottom and top strand). (Provided in Gene editor *in vitro* site-directed mutagenesis system, Table 2.1)

10× annealing buffer. (Provided in Gene editor *in vitro* site-directed mutagenesis system, Table 2.1)

5.2.2.2 Procedure

Half of the denatured pGEM®-11Zf(+) vector (0.5 µg) was used in the hybridisation reaction consisting of: mutagenic oligonucleotide (1 µl), selection oligonucleotide (1 µl), 10× annealing buffer (1 µl) and made up to 20 µl with dH$_2$O, the reaction was carried out as per the manufacturer's instructions.

5.2.3 Mutant strand synthesis and ligation.

The synthesis of the mutant strand takes place in the presence of T4 DNA polymerase, which minimises the risk of secondary undesired mutations and of T4 DNA ligase which ligates the two oligonucleotides with the newly synthesised strand.

5.2.3.1 Materials

10× synthesis buffer. (Provided in Gene editor *in vitro* site-directed mutagenesis system, Table 2.1)

T4 DNA polymerase. (Provided in Gene editor *in vitro* site-directed mutagenesis system, Table 2.1)

T4 DNA ligase. (Provided in Gene editor *in vitro* site-directed mutagenesis system, Table 2.1)

5.2.3.2 Procedure

To the hybridisation mixture, 10× synthesis buffer (3 µl), T4 DNA polymerase (1 µl) and T4 DNA ligase were added and the reaction made up to 30 µl with dH$_2$O. The reaction was carried out at 37°C for 90 min.

5.2.4 Transformation of BMH 71-18 *mut*S and competent *E. coli* JM 109 cells

Transformation is carried out using two different strains of *E. coli* i.e. BMH71-18 *mut*s and JM 109. The former strain enhances the efficiency of mutagenesis at the initial transformation. Transformation in the JM 109 strain ensures separation of the mutant and the non-mutants, consequently giving a high proportion of the mutants.

5.2.4.1 Materials

BMH 71-18 *mut*S and competent *E. coli* JM 109 cells. (Provided in Gene editor *in vitro* site-directed mutagenesis system, Table 2.1)

2 × YT- ampicillin medium. (Section 2.8.1.1)

2 × YT- ampicillin agar plates. (Section 2.8.1.1)

GeneEditor™ Antibiotic selection mix. (Provided in Gene editor *in vitro* site-directed mutagenesis system, Table 2.1)

5.2.4.2 Procedure

Mutant synthesis mixture (2 µl) was used to transform 100 µl of BMH 71-68 *mut*S cells as described by the manufacturer. The plasmid was purified using Wizard® *Plus* SV Minipreps DNA purification protocol and transformed into competent *E. coli* JM 109 cells as per the manufacturer's protocol.

5.2.5 Cutting of pGEM®-11Zf(+) with *Xba*I and *Eco*RI

The pGEM®-11Zf(+) cloning vector was cut with restriction endonucleases *Xba*I and *Eco*RI, so that the fragment constituted by the two annealed primers containing the *Mlu*I restriction site can be inserted by directional cloning. The map of pGEM®-11Zf(+) cloning vector is shown in Figure 5.1

5.2.5.1 Materials
 *Xba*I (Table 2.2)

*Eco*RI (Table 2.2)

10× buffer H. (Provided together with *Eco*RI, Table 2.2)

pGEM®-11Zf(+). (Provided in Gene editor *in vitro* site-directed mutagenesis system, Table 2.1).

5.2.5.2 Procedure
pGEM®-11Zf(+) (2 µg) was cut by *Xba*I (2 µl) in a reaction containing 10× buffer D (5µl), 100 mM BSA (0.5 µl) and made up to 50 µl with dH$_2$O. The digest was carried out overnight at 37°C. *Eco*RI (3 µl) was added to the digest mixture and incubated overnight at 37°C. The *Eco*RI was inactivated by heating the digest mixture in a water bath at 80°C for 30 min.
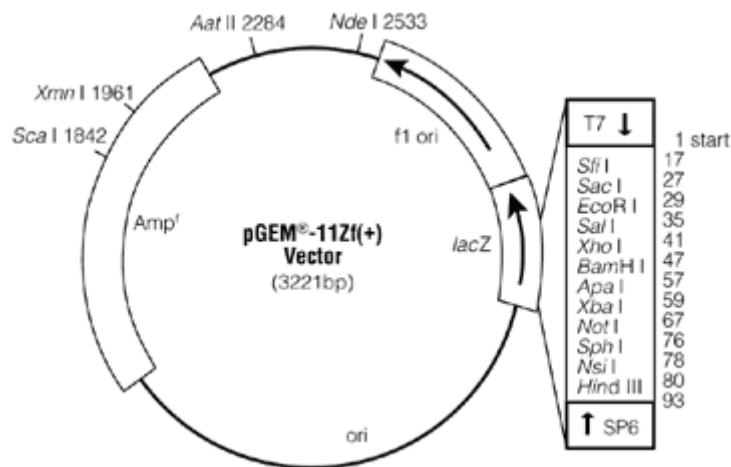
Figure 5.1 Map of pGEM®-11Zf(+) cloning vector www.Promega.com accessed on 12th September 2006.

5.2.6 Dephosporylation of pGEM®-11Zf(+) cut by *Xba*I and *Eco*RI

Dephosphorylation is the process of removing phosphate groups from an organic compound by hydrolysis. Shrimp alkaline phosphatase (SAP) catalyzes the dephosphorylation of 5'

phosphates from DNA and RNA. SAP can be inactivated by heat treatment for 15 min at 65°C unlike the commonly used calf intestinal phosphatases which requires phenol extraction.

5.2.6.1 Materials

Shrimp alkaline phosphatase [1 unit/µl], Table 2.2)

10× dephosphorylation buffer. (Provided with SAP, Table 2.2)

*Xba*I and *Eco*RI cut pGEM 11Z f (+) (Section 5.2.5.2)

5.2.6.2 Procedure

The digestion mix (90 µl) was dephosphorylated in a reaction containing 10× dephosphorylation buffer (20 µl), SAP (3 µl) and made up to 200 µl with dH$_2$O. The reaction was performed at 37°C for 1 h and the SAP was inactivated in a water bath at 65°C for 30 min. Cut and dephosphorylated pGEM®-11Zf(+) was purified using the Wizard® DNA Clean-Up System as described by the manufacturer.

5.2.7 Design of primers containing an *Mlu*I restriction site

For the synthesis of an artificial double stranded DNA fragment containing a *Mlu*I site, two complimentary oligonucleotides were designed, generating once annealed protruding ends corresponding to those produced by restriction with *Eco*RI and *Xba*I. In addition, the primers introduced a *Not*I site, an eight-cutter that has become very common in cloning, and a second *Eco*RI site on the other side of the *Mlu*I site. In the new pMlu vector, the *Mlu*I site is hence flanked by two *Eco*RI sites. We added this feature to our pMlu vector so that any insert cloned on the *Mlu*I site can be excised using *Eco*RI. The rationale behind that is the high cost of *Mlu*I enzyme as compared to *Eco*RI. The first primer sequence was 5'-AATTC GCGGCCGC **ACGCGT** GAATTC T-3' introduced partial *Eco*RI, *Not*I, **Mlu**I, *Eco*RI and partial *Xba*I restriction sites (only a T for the latter) (Figure 5.3) while the second primer was a reverse complement of the first primer with sequence 5'-CTAGA GAATTC **ACGCGT** GCGGCCGC G-3 that introduced partial *Xba*I, *Eco*RI, **Mlu**I, *Not*I and partial *Eco*RI (only a G for the latter) (Figure 5.3). Cutting of the MCS of pGEM®-11Zf(+) with both enzymes removed *Sal*I, *Xho*I, *Bam* HI and *Apa*I sites (Figure 5.2).

The mutagenic primer for site-directed mutagenesis (5'-GCC GAG CTC GA<u>A **CGC GT**</u>C GAC CTCGAG G-3') was homologous to the region of the MCS of pGEM®-11Zf(+) that

encompass restriction sites *Sac*I, *Eco*RI, *Sal*I and *Xho*I, mutating the *Eco*RI site into a *Mlu*I (underlined) site by changing two TT nucleotides into CG (in bold).
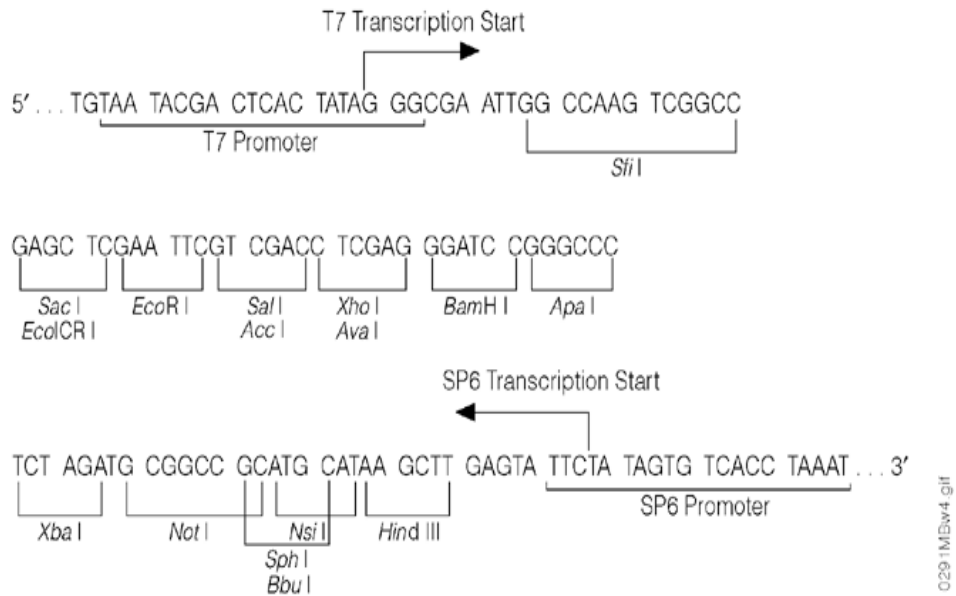
Figure 5.2 **The multiple cloning sites of pGEM®-11Zf(+) cloning vector** www.Promega.com accessed on 12th September 2006.
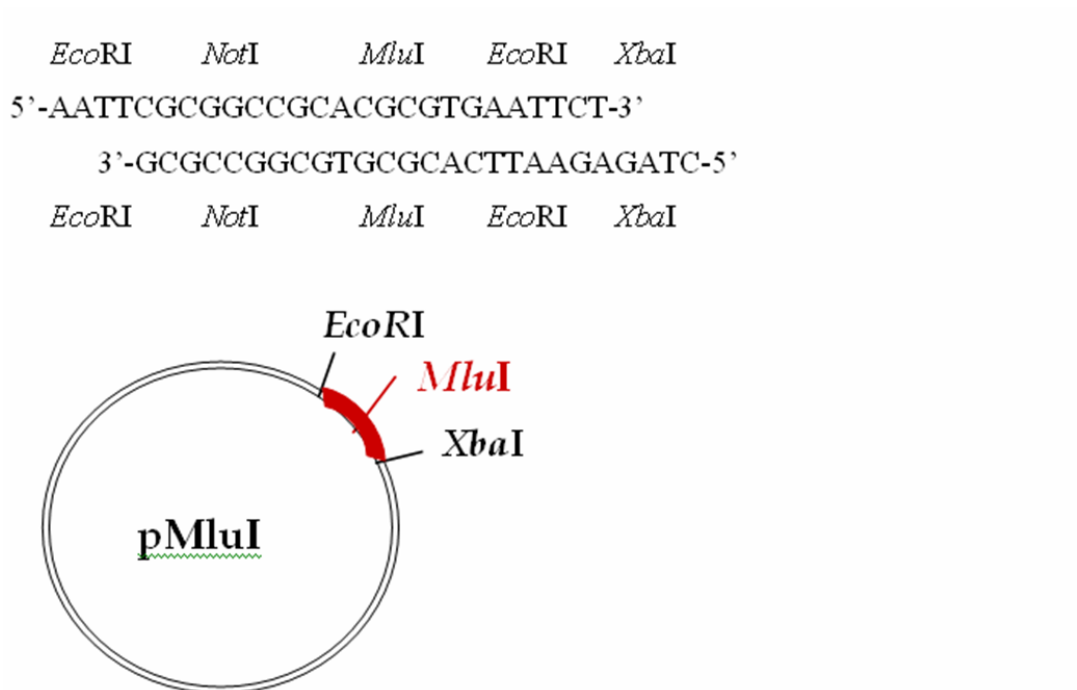
Figure 5.3 Multiple cloning site of pMluI vector. *Eco*RI and *Xba*I, restriction sites used to clone a double stranded fragment incorporating an *Mlu*I restriction site.

5.2.8 Phosphorylation of primers containing an *Mlu*I restriction site

Phosphorylation is the addition of a phosphate group into a protein or a small molecule. T4 polynucleotide kinase (T4 PNK) is a polynucleotide 5'-hydroxyl kinase that catalyzes the transfer of the γ-phosphate from ATP to the 5'-OH group of single- and double-stranded DNA and RNA, oligonucleotides or nucleoside 3'-monophosphates (Richardson, 1965). In the presence of ADP, T4 PNK exhibits 5'-phosphatase activity and catalyzes the exchange of terminal 5'-phosphate groups.

5.2.8.1 Materials
T4 Polynucleotide Kinase [30 U/µl] (Table 2.2)

10× polynucleotide kinase buffer. (Provided with T4 Polynucleotide Kinase, Table 2.2)

F1 primer: AATTCGCGGCCGCACGCGTGAATTC, T_m (67.1°C)

F2 primer: CTAGAGAATTCACGCGTGCGGCCGCG, T_m (67.1°C)

F3 mutagenic primer (Section 5.2.2.1)

100 mM ATP. (Section 2.2)

5.2.8.2 Procedure
The mutagenic primer and both forward and reverse primers (section 5.2.7) were phosphorylated at 37°C for 1 h. The reaction components were: 10× polynucleotide kinase buffer (2.5 µl), T4 DNA polynucleotide kinase (15 units) (0.5 µl), 100 mM ATP (2.5 µl) made up to 25 µl with dH₂O. The T4 DNA polynucleotide kinase was inactivated by heating at 70°C in a water bath for 15 min.

5.2.9 Cloning of a fragment containing an *Mlu*I restriction site into cut and dephosphorylated pGEM®-11Zf(+)

5.2.9.1 Materials
Phosphorylation mix (Section 5.2.8)

Cut and dephosphorylated pGEM®-11Zf(+) (Section 5.6)

5× Rapid ligation buffer (Table 2.1)

T4 DNA ligase (Provided with 5× Rapid ligation buffer, Table 2.1)

5.2.9.2 Procedure
The primers still in the phosphorylation mix were mixed 1:1 with one another and allowed to anneal by cooling the reaction mixture slowly in water bath to obtain a double stranded fragment that can be cloned into pGEM®-11Zf(+). Phosphorylation mix (5 µl) was used in a

ligation reaction consisting of 100 ng (2 µl) of the cut and dephosphorylated pGEM$^®$-11Zf(+), 5× Rapid ligation buffer (4 µl) and T4 DNA ligase (1 µl) made up to 20 µl with dH$_2$O. The ligation was done overnight at room temperature. The ligated vector was transformed into competent *E. coli* JM 109 cells, plasmids containing the *Mlu*I restriction site insert were purified and linearised by *Mlu*I to confirm recombinants, and named pMluI.

5.3   Preparation of pMluI vector for ligation with 2 kb fragments

5.3.1   Materials
pMluI (Section 5.9)

*Mlu*I (Table 2.2)

10 × buffer R (Provided with *Mlu*I, Table 2.2)

100 mM BSA (Section 2.2)

5.3.2   Procedure
pMluI (5 µg) vector was cut with *Mlu*I (20 units) in a 150 µl reaction containing 10 × buffer R (15 µl), 100 mM BSA (1.5 µl) and dH$_2$O (31.5 µl), overnight at 37°C. The vector was purified as described in Wizard$^®$ DNA Clean-Up protocol and dephosphorylated by SAP and separated on a 1% agarose gel stained with ethidium bromide. Cut and dephosphorylated pMluI was excised from the gel and purified as per Wizard$^®$ SV Gel and PCR clean-up protocol.

5.4   Isolation of genomic DNA from trypanosomes
Genomic DNA was isolated from *T. congolense* IL 3000 frozen cells using two methods of DNA extraction i.e. phenol/chloroform extraction and the second using the Fermentas genomic DNA purification kit.

5.4.1   Phenol: chloroform genomic DNA extraction
Isolation of DNA by this method is based on digestion of proteins by Proteinase K and RNA by RNase. After digestion of RNA, RNase and cellular proteins are extracted into the organic phase of buffer saturated phenol:chloroform while the DNA remains in the aqueous phase. Two more phenol:chloroform extractions are done to remove traces of proteins from the DNA. The genomic DNA is precipitated by two volumes of cold ethanol 100% (v/v), washed in 70% (v/v) ethanol and dissolved in Tris-EDTA (TE) or nuclease free water.

5.4.1.1 Materials

Phenol : chloroform : isoamylalcohol (25:24:1). Phenol (100 g) was dissolved in 50 mM Tris-HCl, pH 8 (100 ml) with gentle shaking. The supernatant was removed with a pipette, another 50 mM Tris-HCl, pH 8 (100 ml), added, stirred, and left to stand and form phases. This step was repeated five times to ensure the pH range was between 7 and 8. Chloroform (96 ml) was mixed with isoamylalcohol (4 ml), this mixture was added to the phenol and overlaid with 50 mM Tris-HCl, pH (100 ml) and kept at -20°C (prepared in a fume hood).

500 mM Tris-HCl, pH 8. Tris-base (60.5 g) was dissolved in dH_2O (900 ml), the pH was adjusted to 8 with HCl, made up to 1000 ml with dH_2O and autoclaved (121°C, 30 min, RT)

50 mM Tris-HCl, pH 8. Tris-HCl (500 mM) pH 8 (100 ml) was made up to 1000 ml with sterile dH_2O and stored at RT.

Proteinase K (Table 2.2) 10 mg/ml

RNases A (Table 2.2) 10 mg/ml

T congolense IL 3000 trypanosomes (5×10^8) (Provided by ILRI, Nairobi, Kenya)

Ethanol 100% (v/v) (Section 2.2)

Ethanol 70% (v/v). Ethanol (70 ml) was made up to 100 ml with sterile dH_2O and stored at RT.

3 M Sodium acetate buffer pH 5.2. Glacial acetic acid (18.015/1.05) ml was mixed with dH_2O (90 ml), the pH was adjusted to 5.2 with NaOH and the volume made up to 100 ml with dH_2O.

5.4.1.2 Procedure

Trypanosomes were resuspended in dH_2O (200 µl), RNase A (5 µl) was added before incubation at 37°C for 10 min. Proteinase K was added to a final concentration of 100 µg/ml and incubated at 50°C for 3 h. The DNA was extracted by adding phenol:chloroform:isoamyl alcohol (25:24:1) (400 µl) and gently mixing until an emulsion was formed. The two phases were separated by centrifugation (10,000×*g*, 15 min, RT). The upper aqueous phase containing the genomic DNA was removed and transferred into a new microfuge tube. The DNA was again extracted by adding saturated phenol:chloroform:isoamyl alcohol solution (400 µl) to the aqueous phase, mixing gently and the DNA recovered by centrifugation (10,000×*g*, 15 min, RT). The aqueous phase was removed and transferred into a new microfuge tube. The DNA solution (300 µl) was precipitated by adding 3 M sodium acetate buffer (pH 5.2) (30 µl) and cold 100% (v/v) ethanol (660 µl). The mixture was swirled until a DNA pellet formed, which was removed from the ethanol phase with a very thin glass rod

made from stretching a Pasteur pipette over a gas burner and sealing the end. The pellet stuck to the glass rod was washed in cold 70% (v/v) ethanol and dried at RT. The pellet was then dissolved in dH$_2$O (200 µl) and stored at -20°C.

5.4.2 Fermentas genomic DNA purification method

Genomic DNA is obtained after the trypanosomes are lysed by high salt concentration (that also denatures cellular proteins) present in the lysis solution, while RNA is degraded by RNases. The DNA is then precipitated by ethanol and dissolved in sterile water.

5.4.2.1 Materials
Lysis solution (Provided in the Fermentas Genomic DNA purification kit, Table 2.1)
10 × Precipitation solution (Provided in the Fermentas Genomic DNA purification kit, Table 2.1)
1.2 M Sodium chloride solution (Provided in the Fermentas Genomic DNA purification kit, Table 2.1)
T. congolense IL 3000 trypanosomes (5 x 10^8) (Provided by ILRI, Nairobi, Kenya)
Chloroform (Section 2.2)
Ethanol 100% (v/v) (Section 2.2)
Ethanol 70% (v/v) (Section 5.4.1.1)
RNases (Table 2.2)

5.4.2.2 Procedure
Trypanosome cells were resuspended in dH$_2$O (200 µl), 400 µl of lysis solution was added and incubated at 65°C for 5 min, genomic DNA was isolated as described by the manufacturer's protocol.

5.5 Preparation and cloning of 2 kb inserts from _T. congolense_ IL 3000 genomic DNA into pMluI

5.5.1 Cutting of _T. congolense_ IL 3000 genomic DNA with _Mlu_I

When constructing a genomic library, the DNA is fragmented into smaller fragments that can easily be cloned. As congopain-like genes have a _Mlu_I restriction site in the intergenic region, cutting the gene with this enzyme gives complete genes units (2 kb).

5.5.1.1 Materials
 T. congolense IL 3000 genomic DNA [400 µg/ml] (Provided by ILRI, Nairobi, Kenya)
_Mlu_I [10 units/µl] (Table 2.2)
10× buffer R (Provided with _Mlu_I, Table 2.2)

5.5.1.2 Procedure

T. congolense IL 3000 (30 µg) was cut in a reaction composed of 10 × buffer R (20 µl), *Mlu*I enzyme (3 µl) and dH$_2$O (67 µl). The digestion was done overnight at 37°C, separated on a 0.8% (w/v) agarose gel and DNA with a size of approximately 2 kb (actually probably 1.7 to 2.3 kbp) as estimated using the position of the molecular weight markers on the gel was excised as it corresponds to the size of the cysteine protease genes. The 2 kb fragment was purified from agarose gel slice using DNA purification Wizard® SV Gel and PCR clean-up system according to the manufacturer's protocol.

5.5.2 Cloning of the 2 kb fragment into *Mlu*I cut pMluI

Cut and dephosphorylated pMluI (1 µl) was ligated with purified 2 kb fragments (16 µl) in 20 µl reaction mixtures consisting of 10 × ligation buffer (2 µl) and T4 DNA ligase (1 µl). The ligation was done overnight at room temperature.

5.5.3 Transformation

5.5.3.1 Materials

One shot® Top 10 competent cells (Table 2.1)

S.O.C medium (Provided together with One shot® Top 10 competent cells, Table 2.1)

X-gal [50 mg/ml] (Section 2.9.1.1)

S.O.C-ampicillin agar plates (Section 2.8.3.1)

Ampicillin [100 µg/ml] (Section 2.8.1.1)

5.5.3.2 Procedure

The ligation mix was centrifuged briefly and kept on ice; the vial of competent cells was thawed on ice. The ligation mix (5 µl) was transformed into One shot® Top10 competent cells as described by the manufacturers of the kit. The transformed cells (100 µl) were plated on S.O.C agar plates containing ampicillin (100 µg/ml) and 50 mg/ml X-gal (10 µl). Plates were incubated overnight at 37°C.

5.6 Screening for 2 kb CP clones using SP6/T7 and CP-specific primers

White colonies were screened by PCR using SP6/T7 primers to confirm the presence of an insert. To differentiate congopain-like cysteine proteases in the subgenomic library from non cysteine proteases, further screening was done by PCR using primers Tc/Tv CP2 and Tc/Tv CP3 (Table 4.1) designed from the central domain of the *T. congolense* as radioactive probing was not an option in our laboratory.

5.6.1 Materials
SP6 and T7 primers (Section 2.10.1)

PCR reagents (Table 2.4)

Tc/Tv CP2 and Tc/Tv CP3 primers (Table 4.1)

Cp seq Fw primer TSTTCAAGCAGARCATGGAGC, T_m (52°C)

Cp seq Rev primer TTATTGGRSCCGCACGAGGC, T_m (62°C)

5.6.2 Procedure
PCR was carried out as described in Section 2.10.2. The PCR products were separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2).

5.7 Restriction digest of the CP clones from the subgenomic library with *Pst*I
As shown on figure 3.2 and discussed in Chapter 3, the presence of a band a 4 kbp in cosmid DNA cut with *Pst*I means that a limited number of gene copies do not have a *Pst*I site; such copies would have been missed from the cosmid gene units library described in Chapter 3. To establish whether gene copies lacking *Pst*I site are present in the selected *Mlu*I 2 kbp gene units, clones were restricted with *Pst*I. The multiple cloning site of pMluI does not have a *Pst*I restriction site. To make results more visual, the 2 kb inserts were amplified by PCR using SP6 and T7 primers and digested with *Pst*I.

5.7.1 Materials
8 CP clones from the library (Section 5.6)

*Pst*I (Table 2.2)

10× buffer H (Provided together with *Pst*I, Table 2.2)

*Sal*I (Table 2.2)

10× buffer D (Provided together with *Sal*I, Table 2.2)

5.7.2 Procedure
PCR product (20 µl) was digested in a reaction consisting of *Pst*I (2 µl), 10× buffer H (4 µl), made up to 40 µl with dH$_2$O and incubated overnight at 37°C. Digestion products (20 µl) were separated on a 1.2% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2).

5.8 Purification of mRNA from *T. congolense* IL 3000 and amplification of cysteine protease isoforms by RT-PCR

5.8.1 Materials
T. congolense IL 3000 parasites (Provided by ILRI, Nairobi, Kenya)

Primers: Forward-all-CP and Rev-C1, Rev-C2 and Rev-C3 primers, cluster primers-C1, -C2 and -C3 (Fw and Rev) (Table 3.1)

Primer Rev-C4 5'-CGAGAGAATGCCCACGAAGAGTG-3', T_m (60°C)

SV Total RNA isolation kit (Table 2.1)

Access RT-PCR kit (Table 2.1)

RNAid Kit (Table 2.1)

5.8.2 Procedure

Total RNA was isolated from *T. congolense* clone IL 3000 parasites using SV Total RNA isolation kit as described by the manufacturer and the mRNA subfraction gel purified according to RNAid Kit protocol. Cysteine protease genes were amplified by RT-PCR (Section 2.5.2) using the primers detailed in Table 3.1 as well as Rev-C4 (Section 5.8.1) and separated on an agarose gel and stained with ethidium bromide (Section 2.3.2).

5.9   Results of construction and screening of subgenomic library

5.9.1   Construction of the pMluI cloning plasmid

Of the two methods used in construction of the pMluI cloning vector, only cloning of annealed primers with the *Mlu*I restriction site was successful and was confirmed by cutting the vector with *Mlu*I. The plasmids which were linearised by *Mlu*I were considered recombinant as they contained the *Mlu*I restriction site (results not shown). The site-directed mutagenesis of the pGEM®-11Zf(+) MCS failed as no recombinants with the desired mutations were found.

5.9.1.1 Cutting of pGEM®-11Zf(+) with *Xba*I and *Eco*RI

The pGEM®-11Zf(+) plasmid was linearised by digestion with either *Xba*I or *Eco*RI in the same buffer H, and digestion products (10 μl) separated on a 1% (w/v) agarose gel (fig. 5.5). This was to make sure that both enzymes are able to cut the plasmid in this buffer. Given that the two restriction sites are close to each other, it would not have been possible to tell whether both enzymes had cut by electrophoretic analysis. The analysis with each enzyme was followed by a double digest in buffer H (not shown). The vector appeared to have been cut by the two enzymes, as a unique band at 3.2 kbp is visible. In the uncut vector lanes, two bands are present, at 2 kbp and 4 kbp. They correspond to supercoiled plasmid DNA (covalently closed circular form), monomer and dimer that migrate faster than linear plasmid due to a more compact conformation (1.6 times faster).
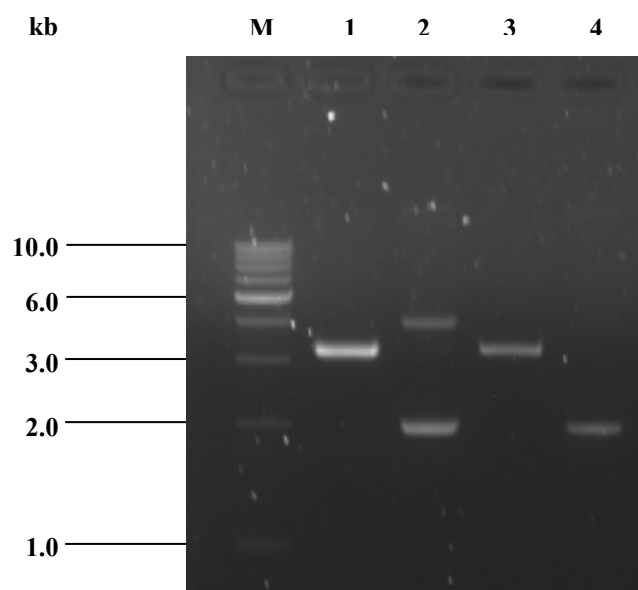
Figure 5.5 Restriction digests of pGEM®-11Zf(+) with *Xba*I and *Eco*RI. The samples were separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2), Lane M, molecular weight makers; lanes 1 and 3, pGEM®-11Zf(+) cut by *Xba*I and *Eco*RI; lanes 2 and 4, uncut pGEM®-11Zf(+).

5.9.2 Preparation of pMluI for ligation with 2 kb fragments.

The pMlu plasmid was restricted with *Mlu*I enzyme and dephosphorylated to prevent recircularisation of the plasmid. The cloning vector appeared linear as shown in Figure 5.6; as explained above, the linearised plasmid migrated slower than uncut plasmid, though no multimeric forms were visible. Cut and dephosphorylated pMluI was gel purified in order to eliminate any traces of uncut plasmid that could be remaining, and could be a source of background (Figure 5.7). This separation of the cut pMluI from the uncut was done in one long well as a large quantity of concentrated cut pMluI was required for subsequent cloning.
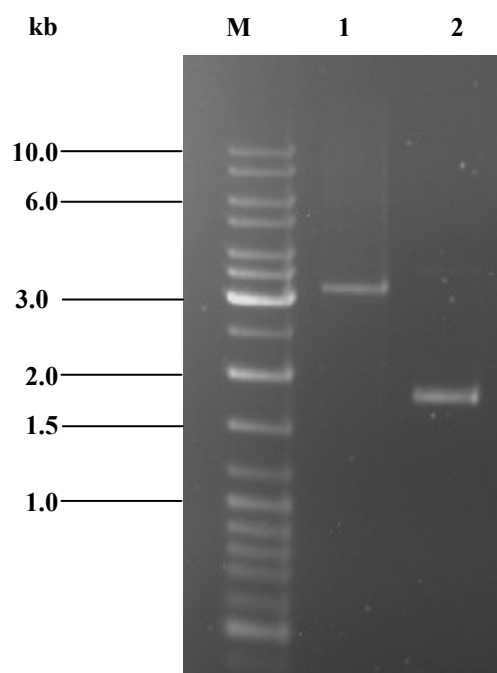
Figure 5.6 Restriction digest of pMluI plasmid with *Mlu*I. Samples were separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Lane M, molecular weight markers; lane 1, pMluI vector cut by *Mlu*I; lane 2, uncut pMluI plasmid.
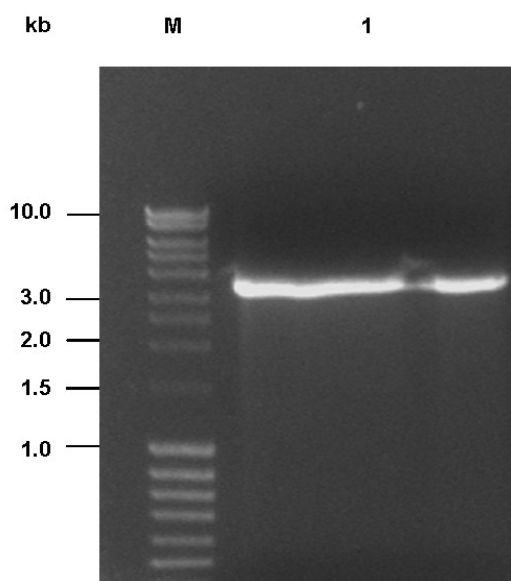
Figure 5.7 pMluI plasmid cut with *Mlu*I and dephosphorylated by SAP. Sample was separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Lane M, molecular weight markers; lane 1, (three teeth of the comb were sealed to make one long well on the gel), pMluI plasmid cut by *Mlu*I and dephosphorylated by SAP.

5.9.3   Isolation of genomic DNA from trypanosomes

Purified *T. congolense* IL 3000 genomic DNA (5 µl) was separated on a 0.8% (w/v) agarose gel to confirm successful isolation from trypanosomes. All the lanes gave two high molecular mass bands showing genomic DNA of good quality that is not degraded. The two bands are likely to correspond to nuclear DNA and to kinetoplast DNA (see section 1.6). The DNA

purified using the kit appears of better quality than the one purified using phenol:chloroform technique.
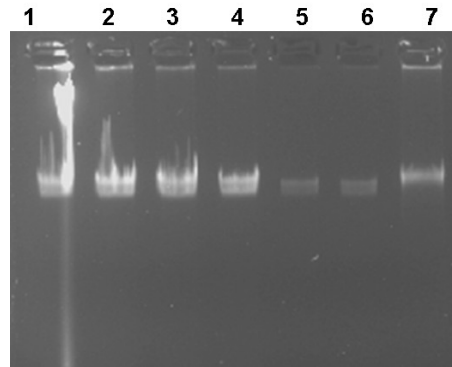
Figure 5.8 Analysis of *T. congolense* IL 3000 genomic DNA isolated from trypanosomes. The *T. congolense* IL 3000 genomic DNA samples (5 μl) were separated on a 0.8% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Lanes 1-6, DNA extraction from frozen trypanosomes using Fermentas Genomic DNA purification method; lane 7, DNA isolated by phenol: chloroform technique.

5.9.4 Cutting of *T. congolense* genomic DNA IL 3000 with *Mlu*I

The genomic DNA was digested by *Mlu*I and separated on a 0.8% (w/v) agarose gel (Figure 5.9). The genomic DNA appeared completely digested by *Mlu*I, a few distinct bands were visualised and the DNA around 2 kb (in the fact, 1.7 to 2.3 kbp) as shown by the arrow Figure 5.9) representing the size of the cysteine protease genes, was excised from the gel.
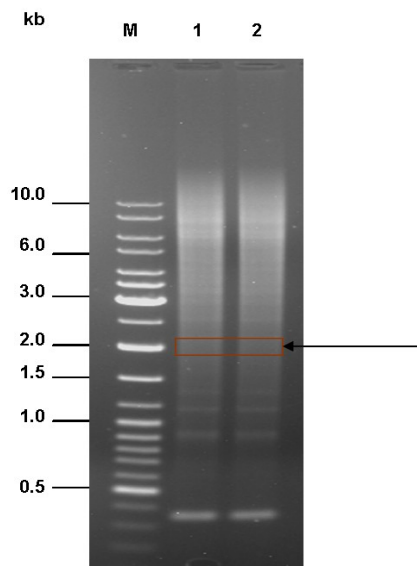
Figure 5.9 Restriction digest of *T. congolense* IL 3000 genomic DNA using *Mlu*I. The DNA samples were separated on a 0.8% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2), lane M, molecular weight markers; lanes 1 and 2, *T. congolense* IL 3000 genomic DNA completely digested by *Mlu*I. The arrow indicates the region on the gel containing the digested genomic DNA of 2 kb that was excised from the gel for further studies.

5.9.5   Cloning of 2 kb fragment into pMluI cut with *Mlu*I

Purified genomic DNA in the 1.7-2.3 kbp range was ligated into pMlu at the *Mlu*I site and the ligation mix used to transformation one shot® Top10 competent cells. Many white colonies were visible on Xgal-containing S.O.C-ampicillin agar plates, with a 10 to 15 % background of blue colonies. White colonies were considered as recombinant and further screened by PCR to establish the clones carrying the 2 kb fragment.

5.9.6   Screening for CP clones using CP-specific primers

White colonies of pMluI/2 kb clones were screened by PCR using SP6/T7 primers to confirm the presence of a 2 kb insert and representative analyses are shown in Figure 5.10 A and B (lanes 1, 9, 16, 18, 20, 21, 22, 23, 27, 30 and 30). A total of 206 recombinants were obtained and only 8 clones (2, 3, 4, 6, 30, 31, 50 and 203) were identified as containing congopain-like cysteine protease by PCR using CP-specific primers, prior to sequencing (Figure 5.11 A, B and C, lanes 11, 30, 31 and 50). The rest of clones were considered non cysteine protease genes because they failed to amplify.

A

B
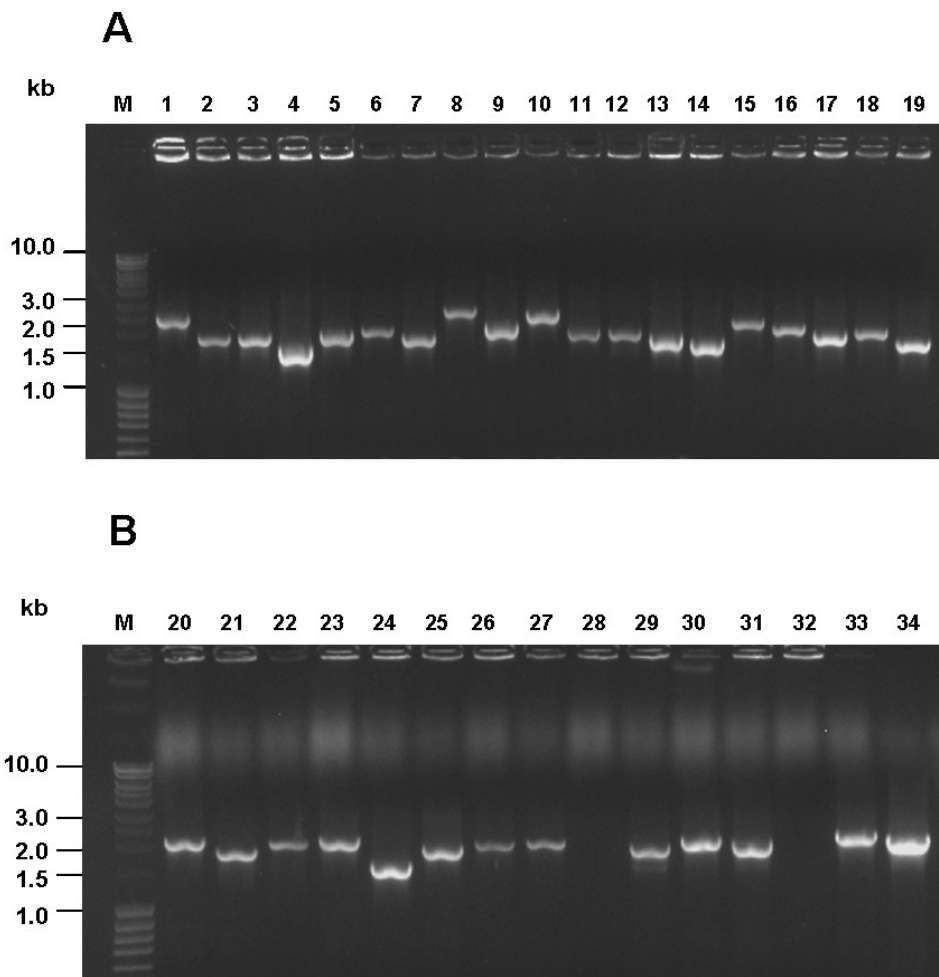
Figure 5.10 PCR on colonies of each clone of the 2 kb library using SP6 and T7 primers. The samples were separated on a 1 % (w/v) agarose gel and stained with ethidium bromide. (Section 2.3.2), **(A)** lane M, molecular weight makers; lanes 1-19, PCR on colonies 1-19 **(B)** lane M, molecular weight makers; lanes 20-32, PCR on colonies 20-32 amplified by SP and T7 and lane 33 and 34 are clones 3 and 6 respectively (positive controls).
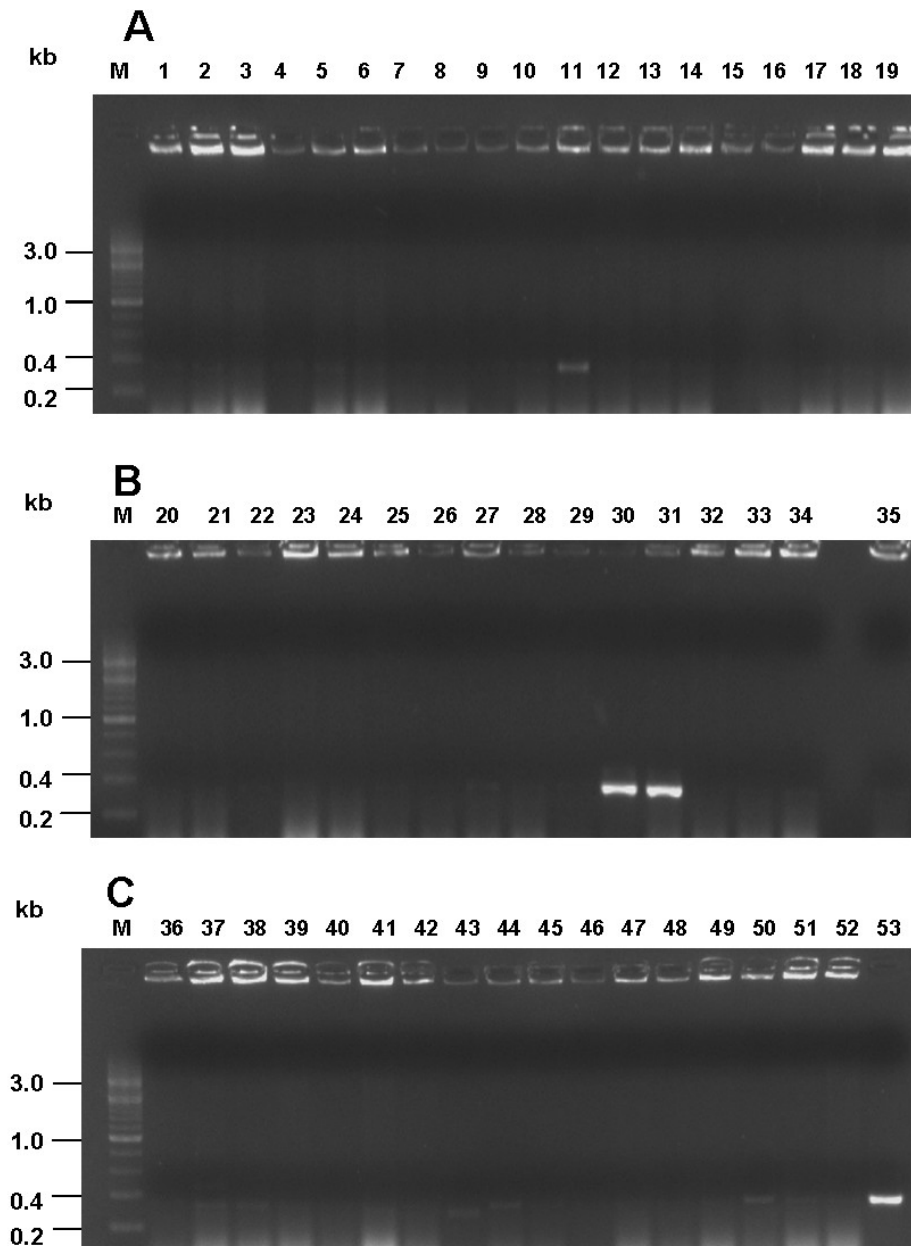
Figure 5.11 PCR on recombinant colonies amplified by CP-specific primers. The samples were separated on a 1.2% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Lanes M, molecular weight markers. **(A)** Lanes 1-19, recombinant colonies 1-19 amplified by CP-specific primers. **(B)** Lanes 20-35, PCR on recombinant colonies 20-35. **(C)** lanes 36-51, recombinant colonies 36-51 amplified by CP-specific primers, lane 52, a negative control (PCR on a blue colony) and lane 53, a positive control (PCR on clones 3).

5.9.7   Restriction mapping of the CP clones from the subgenomic library

The insert of the 8 positive clones were amplified by PCR using primers flanking the MCS (so-called T7 and SP6), and the PCR product cut with *Pst*I. The digests were separated on a 1.2% (w/v) agarose gel containing ethidium bromide. All eight CP positive clones were cut by *Pst*I resulting in two bands at 500 bp and 2 kb (Figure 5.12). Hence the copies that do not have a *Pst*I site that we hoped would be present in the library were not selected.

The NEBcutter V2.0 (www.tools.neb.com/NEBcutter2) was used to establish the restriction map of the clones. All the clones had *Pst*I, *Bam*HI, and *Aat*II among other restriction sites common to all trypanosomal cysteine protease. Interestingly, it was previously thought that the *Sal*I restriction site was confined to CP1 isoforms, while the *Nco*I restriction site was associated with CP2 and CP3 isoforms, however clones 4, 6 and 50 which had both *Sal*I and *Nco*I restriction sites. *Hind*III and *Xba*I restriction sites were also thought not to exist within the CP genes however, clone 30 had an *Xba*I restriction site while clone 203 had a unique *Hind*III restriction site. Clones 30, 31 and 203 were the only clones without an *Nco*I restriction site.
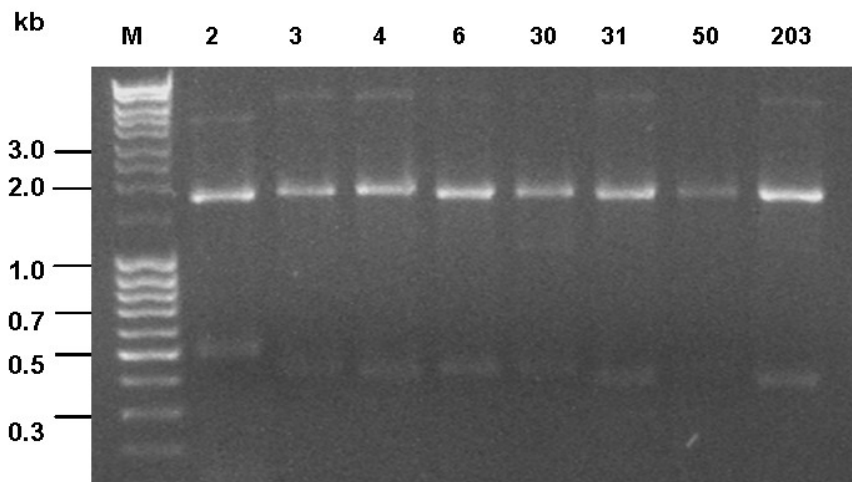
Figure 5.12 *Pst*I and *Sal*I restriction digest of CP clones amplified with SP6 and T7 primers. The samples were separated on a 1.2% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). **(A).** Lane M, molecular weight markers; numbers correspond to the clone numbers digested by *Pst*I. **(B).** Lanes M, molecular weight markers; numbers correspond to the clone numbers digested using *Sal*I.

5.9.8   Sequence analysis

The eight CP clones were sequenced with SP6 and T7 primers (Section 2.10.1) from both ends and internally by CP-seq Fw and Rev primers to get the full sequence. There was a problem in sequencing clone 2 and its sequence was excluded from the analysis. The DNA sequences were compared with other congopain-like cysteine protease genes in available at NCBI BLAST (www.ncbi.nlm.nih.gov/BLAST, accessed on 15.11.2006). At the amino acid level the seven clones showed significant identity (86-94%) with CP1 (Fish *et al*., 1995, EMBL accession number Z25813), CP2 (Jaye *et al*., 1994, EMBL accession number L25130) and CP3 (Downey and Donelson, 1999 EMBL accession no. AF139913). The DNA sequences were translated into protein sequences by ExPASy software and aligned with CP1, CP2 and CP3 by ClustalW, and showed high identity (86-100%).

The signal peptide (20 amino acids long) was 100% identical (Figure 5.13, text box 1); while the propeptide was 95-100% identical and 105 amino acids long (Figure 5.13). The inhibitory peptide YHNGA sequence (Figure 5.13, text box 2) in the propeptide (Lalmanach *et al*., 1998) was present in all the clones. All the clones had the 22 nucleotide polypyrimidine tract (CTTTTCTTCTTTCTTTTCCTTT) that precedes the start codon of cysteine proteases in *T. congolense* (results not shown). The catalytic domains of the clones were 216 amino acid residues long and showed some degree of variability (79-96 % identity). Clones 6 and 50 had identical catalytic domains (Figure 5.14). Clones 4, 6, 30 and 50 had the APEA sequence at the maturation site (N-terminus of mature protease), while clone 203 had APPA, clone 3 had APDA and clone 31 had the PPMT (Figure 5.14, text box 1). The N-terminus sequences APPA and APEA have previously been described in CP1 and CP2 respectively (Boulangé *et al*., 2001). Among the seven clones only clone 4 had a conserved cysteine residue in the catalytic triad, the other clones had a serine residue (Figure 5.14, text box 2). The histidine residue of the catalytic triad was conserved in clones 3, 4, 6, 50 and 203, while in clones 30 and 31 it was substituted by a serine residue (Figure 5.14, text box 3), first reported in CP3 (Downey and Donelson, 1999 EMBL accession no. AF139913). The asparagine residue of the catalytic triad was present in all the clones (Figure 5.14, text box 4).

The length of the C-terminal extension varied among the clones; it was 97 amino acids long in clones 6, 30, 31, 50 and 203, 95 amino acids long in clone 3 and 98 amino acids long in clone 4 (Figure 5.15). The clones had a C-terminal extension linked to the central domain by polyproline residues and were 90 – 92 % identical. Clones 4, 6, 50, 30, 31, 50 and 203 had

six proline residues while clone 3 had 8 proline residues as was the case for CP1 (Figure 5.15, text box 1).

A Neighbouring Joining tree performed on translated amino acid sequences of the seven clones together with cysteine protease isoforms CP1, CP2 and CP3, formed six clusters (Figure 5.16). The first cluster consisted of clones 6 and 50, these clones had a serine residue instead of cysteine at the active site but the other residues forming the catalytic triad were conserved. Clone 3 fell into the second cluster and like clones 6 and 50 had a serine residue at the active site and the histidine and asparagine residues of the catalytic triad were conserved. However in contrast to clones 6 and 50 which had an N-terminal APEA sequence, clone 3 had N-terminal APDA sequence. The third cluster comprised of clone 4 and CP1 both of which had the three amino acid residues of the catalytic triad (CHN) conserved, but differed in their N-terminal sequences: clone 4 (APEA) and CP1 (APPA). Surprisingly, the fourth cluster comprised CP2 and CP3, which differ in their amino acid residues at the active site of the mature enzyme, CP2 (cysteine) and CP3 (serine). CP2 and CP3 also differed in their N-terminal sequences (APEA for CP2 and APPA for CP3). The fifth cluster comprised of clone 203 occurring between CP3 and clones 30 and 31; this clone had a serine residue instead of a cysteine residue but histidine and asparagine residues conserved. The last cluster (sixth) comprised of clones 30 and 31, bearing a serine residue at the active site, a serine residue instead of histidine and a conserved asparagine residue. Clone 30 had the APEA N-terminus while clone 31 had the unique PPMT sequence. The seven clones appeared closely related to CP1 and CP3 than CP2, clone 4 appeared closest to CP1 with 93% identity.

```
CLUSTAL 2.0.3 multiple sequence alignment

                              1
Clone3      MPRSEMTRTLRFSVGLLAVA ACFVPVALGVLHAEQSLQQQFAAFKQKYSRSYKDATEEAF 60
Clone4      MPRSEMTRTLRFSVGLLAVA ACFVPVALGVLHAEQSLQQQFAAFKQKYSRSYKDATEEAF 60
Clone6      MPRSEMTRTLRFSVGLLAVA ACFVPVALGVLHAEQSLQQQFAAFKQKYSRSYKDATEEAF 60
Clone31     MPRSEMTRTLRFSVGLLAVA ACFVPVALGVLHAEQSLQQQFAAFKQKYSRSYKDATEEAF 60
Clone50     MPRSEMTRTLRFSVGLLAVA ACFVPVALGVLHAEQSLQQQFAAFKQKYSRSYKDATEEAF 60
CP2         MPRSEMTRTLRFSVGLLAVA ACFVPVALGVLHAEQSLQQQFAAFKQKYSRSYKDATEEAF 60
CP3         MPRSEMTRTLGFSVGLHAVA ACFVPVALGVLHAEQSLQQQFAAFKQKYSRSYKDATEEAF 60
CP1         MPRSEMTRTLRFSVGLLAVA ACFVPVALGVLHAEQSLQQQFAAFKQKYSRSYKDATEEAF 60
Clone203    MPRSEMTRTLRFSVGLLAVA ACFVPVALGVLHAKQSLQQQFAAFKQKYSRSYKDATEEAF 60
Clone30     MPRSEMTRTLRFSVGLLAVA ACFVPVALGVLHAEQSLQQQFAAFKQKYSRSYRDATEEAF 60
            *********  ***** ***  *************:*******************:*******
                                            2
Clone3      RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKRPRKVVT 120
Clone4      RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKRPRKVVT 120
Clone6      RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKRPRKVVT 120
Clone31     RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKRPRKVVN 120
Clone50     RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKRPRKVVT 120
CP2         RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKRPRKVVN 120
CP3         RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKRPRKVVN 120
CP1         RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKRPRKVVN 120
Clone203    RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEEFRAT YHNGA EYYAAALKRPRKVVN 120
Clone30     RFRVFKQNMERAKEEAAANPYATFGVTRFSDMSPEVRKAT YHNGA EYYAAALKRPRKVVN 120
            ***********************************  :** ***** ***************.

Clone3      VSTGK-- 125
Clone4      VSTGK-- 125
Clone6      VSTGK-- 125
Clone31     VSTGR-- 125
Clone50     VSTGK-- 125
CP2         VSTGK-- 125
CP3         VSTGK-- 125
CP1         VSTGK-- 125
Clone203    VSTGK-- 125
Clone30     VSTGK-- 125
            ****:
```

Figure 5.13 ClustalW alignment of translated amino sequences of subgenomic library clones showing the signal and propeptide regions of congopain-like cysteine proteases. Clones 3, 4, 6, 30, 31, 50 and 203, translated amino acid sequences of clones 3, 4, 6, 30, 31, 50 and 203 respectively; CP1, CP2 and CP3 are translated amino acid sequences of cysteine protease isoforms CP1, CP2 and CP3 respectively. Text box 1, the signal peptide of congopain-like cysteine proteases; text box 2, the YHNGA peptide in the propeptide. The remainder of the amino acid residues after text box 1, represent the 105 amino acid long propeptide region of congopain-like cysteine proteases.

```
CLUSTAL 2.0.3 multiple sequence alignment


           1                    2
Clone6     APEA VDWRKKGAVTPVKDERLCDS S WAFSAIGNIEGQWKVAGHELTSLSEQMLLSCDTRE 60
Clone50    APEA VDWRKKGAVTPVKDERLCDS S WAFSAIGNIEGQWKVAGHELTSLSEQMLLSCDTRE 60
Clone3     APDA VDWRKKGAVTPVRDERLCDS S WAFSAIGNIEGQWKVAGHELTSLSEQMLLSCDTRE 60
Clone4     APEA VDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKVAGHELTSLSEQMLVSCDTED 60
CP1        APPA VDWRKKGAVTPVKDQGACGS C WAFSAIGNIEGQWKVAGHELTSLSEQMLVSCDTTD 60
CP2        APEA VDWRKKGAVTPVKDQGQCGS C WAFSAIGNIEGQWKVAGHELTSLSEQMLVSCDTND 60
CP3        APPA IDWRKKGAVTPVKDQGQCHS S WAFSAIGNIEGQWKIAGHELTSLSEQMLVSCDTND 60
Clone203   APPA IDWRKKGAVTPVKDQGQCGS S WAFSAIGSIEGQWKVAGHELTSLSEQMLVSCDTND 60
Clone31    PPMT VDWRKKGAVTPVKDQGQCDS S WAFSAIGNIEGQWKIAGHELTSLSEQMLVSCDTND 60
Clone30    APEA VDWRNLGAVTPVKDQGKCDS S WAFTVIGNIEGQWKIAGHELTSLSEQMLVSCDTND 60
           * : :*** ****** *:   * *  . * * . ** .****** :************** :**** :


Clone6     DGCGGGLMDRAFKWIVSSNKGNVFTEQSYPYASTDGDVPRCNKSGKVVGAKIRDHVDLPQ 120
Clone50    DGCGGGLMDRAFKWIVSSNKGNVFTEQSYPYASTDGDVPRCNKSGKVVGAKIRDHVDLPQ 120
Clone3     DGCGGGLMDRAFQWIVSSNKGNVFTEQSYPYASTDGDVPRCNKSGKVVGAKISDYVDLPQ 120
Clone4     DGCGGGLMDNAFQWIVSSNKGNVFTEQSYPYASKGGKMPPCNMSGKVVGANIRDHVDLPK 120
CP1        YGCRGGLMDKSLQWIVSSNKGNVFTAQSYPYASGGGKMPPCNKSGKVVGAKISGHINLPK 120
CP2        FGCEGGLMDDAFKWIVSSNKGNVFTEQSYPYASGGGNVPTCDKSGKVVGAKIRDHVDLPE 120
CP3        FGCGGGFSDPAFKWIVSSNKGNVFTEQSYPYASGGGNVPTCDKSGKVVGAKIRDRVDLPR 120
Clone203   LGCKLGLMDFAFKWILWSNKGNVFTEQSYPYASGGGKVPTCNKSGKVVGAKISDWIHFPR 120
Clone31    FGCGGGFSDPAFQWILSSNKGNVFTEQSYPYASGGGNVPACNMSGKVVGAKISNRLYLPR 120
Clone30    FGCRAGFMDTAFKWIVSPNDGNVFTEQSYPYASGGGNVPACNKSGKVVGANIRDHVHILD 120
            **   *: *  ::::**: .*.***** ******* .*.:* *: *******:* . : :


                                                    3
Clone6     DENAIAEWLAKNGPVAIAVDATSLQRYTDGVLTSCISEQLD H GVLLVGYDDTSKPPYWII 180
Clone50    DENAIAEWLAKNGPVAIAVDATSLQRYTDGVLTSCISEQLD H GVLLVGYDDTSKPPYWII 180
Clone3     DENAIAEWLAKNGPVAIAVEATSLQRYTGGVLTSCISEQLD H GVLLVGYDDTSKPPYWII 180
Clone4     DENAIAEWLAKNGPVAIAVDATSFLDYTGGVLTSCISKQLD H GVLLVGYDDTSKPPYWII 180
CP1        DENAIAEWLAKNGPVAIAVDATSFLGYKGGVLTSCISKGLD H DVLLVGYDDTSKPPYWII 180
CP2        DENAIAEWLAKNGPVAIAVDATSFQSYTGGVLTSCISEHLD H GVLLVGYDDTSKPPYWII 180
CP3        DENAIAEWLAKKGPVAIAVDATSFQSYTGGVLTSCISEHLD H GVLLVGYDDTSKPPYWII 180
Clone203   DEDIIAEWLAKNGPVAIAVDSTSFQSYTGGVLTSCISEYLD H AVLLVGYDDTSKPPYWII 180
Clone31    DEDMIAEWLAKKGPVAIAVDATSFQSYTGGVLTSCISKEVN S AALLVGYDDTSKPPYWII 180
Clone30    NENAIAEWLAKNGPVAIAVDATSFQRYTGGVLTSCISKEVN S AALLVGYDDTSKPPYWII 180
           :*: *******:*******::**:  *..********: ::    .***************


           4
Clone6     K N SWGKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG-- 216
Clone50    K N SWGKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG-- 216
Clone3     K N SWGKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG-- 216
Clone4     K N SWSKGWGEEGYIRIEKGTNQCLMKNYASSAVVSG-- 216
CP1        K N SWSKGWGEEGYIRIEKGTNQCLMKNYARSAVVSG-- 216
CP2        K N SWSKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG-- 216
CP3        K N SWGKGWGEEGYIRIEKGTNQCLMKNLPSSAVVSG-- 216
Clone203   K N SWSKGWGEEGYIRIEKGTNQCLVQEYASSAVVIG-- 216
Clone31    K N SWSKGWGEEGYIRIEKGTNQCLVQEYASSAVVSG-- 216
Clone30    K N SWGKGWGEEGYIRIEKGTNQCRMKDYVSSAVVSG-- 216
           *  * **.***************** :::    **** *
```

Figure 5.14 ClustalW alignment of translated amino sequences of subgenomic library clones showing the catalytic domain of congopain-like cysteine proteases. Clones 3, 4, 6, 30, 31, 50 and 203, translated amino acid sequences of clones 3, 4, 6, 30, 31, 50 and 203 respectively; CP1, CP2 and CP3 are translated amino acid sequence of cysteine protease isoforms CP1, CP2 and CP3 respectively. Text box 1, cleavage site between the propeptide and the catalytic domain of congopain-like cysteine proteases; text box 2, cysteine or serine residues at the active site; text box 3, histidine or serine residues of the catalytic triad; text box 4, asparagine residue of the catalytic triad.

```
CLUSTAL 2.0.3 multiple sequence alignment

                       1
Clone6      PVPPPPPPAS TFTQEFCEGAEC-SGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 59
Clone50     PVPPPPPPAS TFTQEFCEGAEC-SGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 59
Clone4      PVPPPPPPAS TFTQEFCEGAECQSGCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 60
Clone3      PVPPPPPPPP TFTQEFCEGAQC-SGCTKATFPTGKCVQLSGAGSVIASCGPNNLTQIVYP 59
CP2         --PPPPPPTP TFTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
CP3         --PPPPPPTP TFTQELCEGAECQSKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 58
Clone31     --PPPPPPTP TFTQELCEGAEC-SKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 57
Clone30     --PPPPPPTP TFTQELCEGAEC-SKCTKATFPTGKCVQLSGAGSAIASCGSNNLTQIVYP 57
Clone203    --PPPPPPTP TSTQELCEGAEC-SKCTKATFPTGKCVQLSGAGSVIASCGSNNLTQIVYP 57
CP1         PPPPPPPPAS TFTQEFCEGAECQSGCTKATFPTGKCVQFGGAGSVIASCGSNNLTQIVYP 60
                 ******.. *  ***:****:*  *  **************:.****.*****.*********

Clone6      LSSSCSGYSVPLTVPLDKCLPIVIGSVMYECSGKAPTESARLVRHE--- 105
Clone50     LSSSCSGYSVPLTVPLDKCLPIVIGSVMYECSGKAPTESARLVRHE--- 105
Clone4      LSSSCSGYSVPLTVPLDKCLPIVIGSVMYECSGKAPTESARLVRHE--- 106
Clone3      LSSSCSGYSVPLTVPLDKCLPIVIGSVMYECSGKAPTESARLVRHE--- 105
CP2         LSSSCSGFSVPLTVPLDKCLPIVIGSVMYECSDKAPTESARLVRHE--- 104
CP3         LSSSCSGFSVPLTVPLDKCLPIVIGSVMYECSDKAPTDPPG------- 99
Clone31     LSSSCSGFSVPLTVPLDKCLPIVIGSVMYECSDKAPTESSRLVRHE--- 103
Clone30     LSSSCSGFSVPLTVPLDKCLPIVIGSVMYKCSDKAPTASARLVRHE--- 103
Clone203    LSSSCSGFSIPLTVPLDKCLPIVIGSVMYECSDRAPTESARLVRHE--- 103
CP1         LSSSCSGFSIPLTVPLDKCLPIVVGSVMYECSGKAPTESARLVRHE--- 106
                 *******:*:**************:*****:**.:***  ..
```
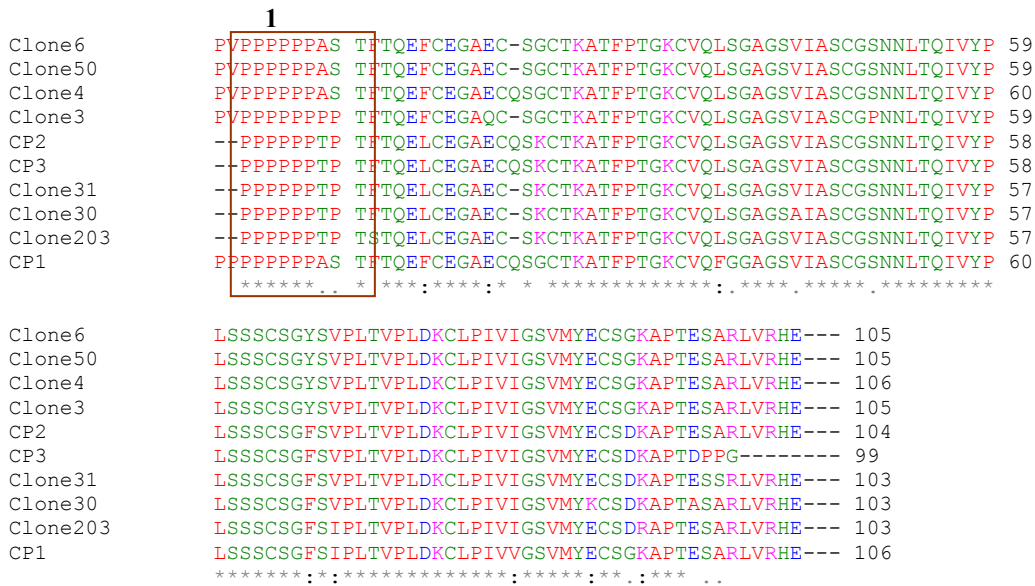
Figure 5.15 ClustalW alignment of translated amino sequences of subgenomic library clones showing the polyproline hinge and the C-terminal extension of congopain-like cysteine proteases. Clones 3, 4, 6, 30, 31, 50 and 203, translated amino acid sequences of clones 3, 4, 6, 30, 31, 50 and 203 respectively; CP1, CP2 and CP3 are translated amino acid sequence of cysteine protease isoforms CP1, CP2 and CP3 respectively. Text box 1, the polyproline hinge linking the catalytic domain to the C-terminal extension of congopain-like cysteine proteases.
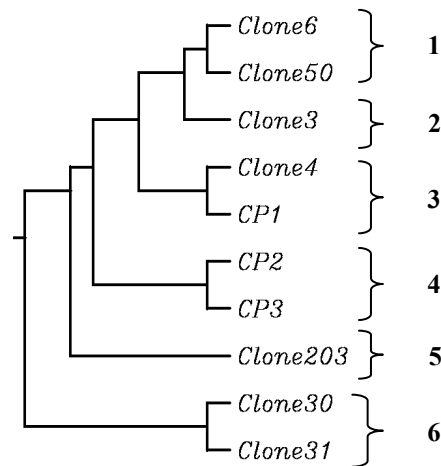

Figure 5.16 Neighbouring Joining tree performed by ClustalW alignment on translated amino acid sequence of entire congopain-like cysteine protease genes (clones) obtained from the subgenomic library. CP1, CP2 and CP3 are cysteine protease isoforms CP1, CP2 and CP3 respectively. Clones 3, 4, 6, 30, 31, 50, and 203, predicted amino acid sequences of the corresponding clones used in the construction of the Neighbouring Joining tree. Numbers 1-6, clusters formed upon alignment.

5.10   Results of detection of cysteine protease isoform transcript by RT-PCR

5.10.1 Isolation of total *T. congolense* IL3000 RNA

Total RNA was successfully isolated from *T. congolense* IL 3000 trypanosomes as shown in Figure 5.17. Since the open reading frames of CP1 and CP2 are 1334 bp and 1335 bp respectively and also considering the length of the poly $(A)^+$ tail, the RNA around 1.5-1.7 kb was excised and purified to increase RT-PCR efficiency. The eukaryotic (5S, 5.8S, 18S and 28S) and prokaryotic (5S, 16S and 23S) rRNA subunits, were visible on the gel (Figure 5.17). Prokaryotic rRNA originates from the kinetoplast (a large mitochondrion).
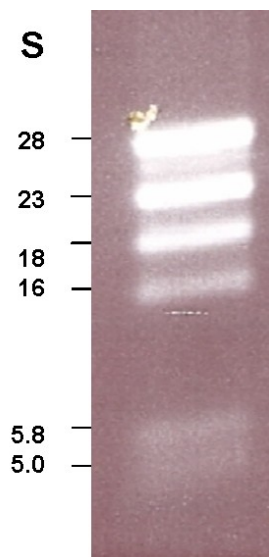

Figure 5.17 Total RNA isolated from *T. congolense* IL 3000 parasites. Total RNA was separated on a 1% (w/v) agarose gel and stained with ethidium bromide (Section 2.3.2). Bands represent the subunits of eukaryotic and prokaryotic rRNA.

5.10.2 RT-PCR of cysteine protease isoforms mRNA

The mRNA transcript of the different cysteine protease isoforms were targeted by RT-PCR using cluster-C1, -C2, -C3 (Fw and Rev) and another set of three primers with a common forward primer (Forward-all-CP) and reverse primers (Rev-C1,-C2,-C3, -C4; Table 3.1). All the primer sets gave a PCR product of the expected sizes (Figure 5.18). Cluster primers-C1, -C2 and -C3 (Fw and Rev) amplified as expected a fragment of approximately 250 bp (lanes 5-7). The common forward primer (Forward-all-CP) and the reverse primers gave a band at ~ 700 bp (C1-Rev), ~ 400 bp for C2-Rev, ~ 850 bp for C3-Rev and ~ 300 bp for C4-Rev as shown in lanes 1-4 (Figure 5.18). These results suggested that all the CP families (CP1, CP2 and CP3) were expressed in bloodstream forms of *T. congolense*.
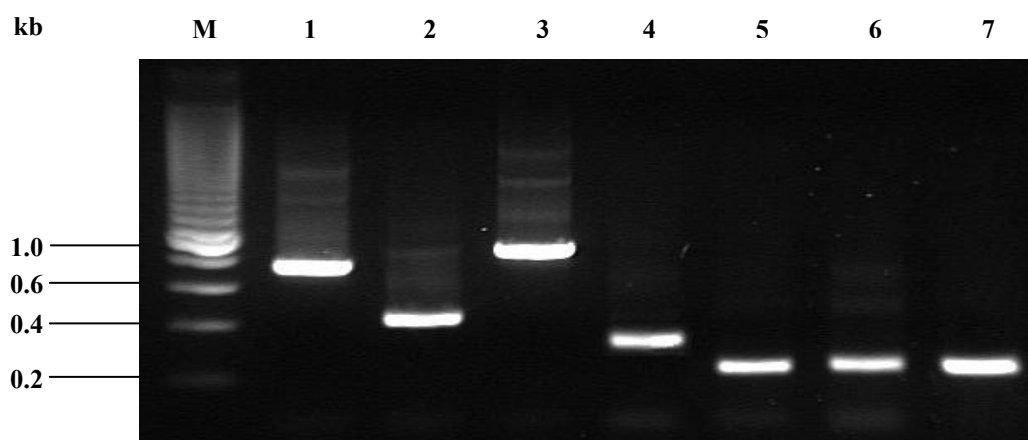
Figure 5.18 RT-PCR on *T. congolense* IL 3000 congopain-like cysteine protease genes. The samples were separated on a 1% (w/v) agarose gel stained with ethidium bromide (Section 2.3.2). Lane M, molecular weight markers; lane 1, mRNA amplified by primers Forward-all-CP and Rev-C1 (CP1); lane 2, RT-PCR using Forward-all-CP and Rev-C2 (CP2); lane 3, RT-PCR by Forward-all-CP and Rev-C3 ; lane 4, mRNA amplified by Forward-all-CP and Rev-C4; lane 5, RT-PCR using cluster primers-C1 (Fw and Rev); lane 6, RT-PCR using cluster primers-C2 (Fw and Rev) and lane 7, RT-PCR by cluster primers-C3 Fw and Rev.

5.10.3 Control experiment using *Taq* DNA polymerase

RNA molecules are much more fragile than DNA, hence RNA preparation are often contaminated by traces of genomic DNA. If for most applications this is not a problem, for RT-PCR, it could generate false positive. It is therefore important to check, by PCR, that no DNA is present in RNA samples that could be amplified by a thermostable polymerase. A negative control experiment in which *Taq* DNA polymerase was used for RT-PCR instead of *Tfl* DNA polymerase failed to show a PCR product (Figure 5.19). This confirmed that there was no contaminating DNA in the mRNA preparation used to conduct the reverse transcription. A positive control, *T. congolense* IL3000 genomic DNA was amplified by *Taq* DNA polymerase using primers Forward-all-CP and Rev-C2 (CP2) and Rev-C1 (CP1) resulting in bands of the expected size of ~ 400 bp and ~ 700 bp respectively (Figure 5.19), identical to those obtained from mRNA using *Tfl* DNA polymerase (Figure 5.18).
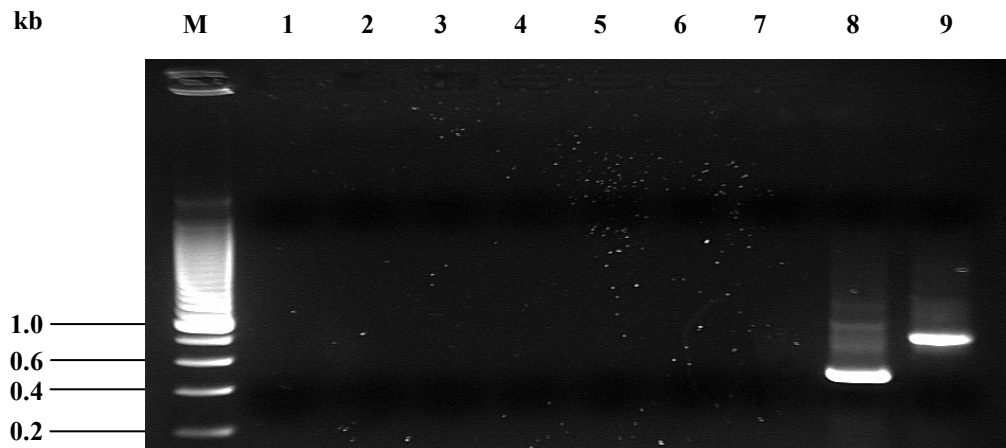
Figure 5.19 RT-PCR on *T. congolense* IL 3000 congopain-like cysteine protease using *Taq* polymerase DNA. Samples were separated on a 1% (w/v) agarose gel stained with ethidium bromide (Section 2.3.2). Lane M, molecular weight markers; lane 1, mRNA amplified by primers Forward-all-CP and Rev-C1 (CP1); lane 2, RT-PCR using Forward-all-CP and Rev-C2 (CP2); lane 3, RT-PCR by Forward-all-CP and Rev-C3 ; lane 4, mRNA amplified by Forward-all-CP and Rev-C4 ; lane 5, RT-PCR using cluster primers-C1 (Fw and Rv); lane 6, RT-PCR using cluster primers-C2 (Fw and Rev); lane 7, RT-PCR by cluster-C3 (Fw and Rev) primers; lane 8, PCR on *T. congolense* IL 3000 genomic DNA using Forward-all-CP and Rev-C2 (CP2) and lane 9, PCR on *T. congolense* IL 3000 genomic DNA using Forward-all-CP and Rev-C1.

5.11   Discussion

In a previous study, cysteine protease genes were found to exist at a single locus of approximately 25 kb and in a tandem repeat of 12-14 copies in *T. congolense* strain IL3000 (Boulangé et al., manuscript in preparation). Similarly, a multi-gene family of cysteine protease genes has been reported having 20 copies in *T. brucei* (Mottram *et al*., 1989), 19 copies in *L. mexicana* (Souza *et al*., 1992), 6 copies in *T. cruzi* (Eakin *et al*., 1992), eight SERA-like genes in *P. falciparum* (Gardner *et al*., 2002), five SERA-like genes in *P. vivax* (Kiefer *et al*., 1996) and three SERA homologues in *P. vinckei* (Gor *et al*., 1998). However, other cysteine protease genes such as cathepsin L-like *cpa* exists as single copy in *L. mexicana* (Mottram *et al*., 1998), *L. pifanoi* (Traub-Cseko *et al*., 1993) and in *L. donovani chagasi* (Omara-Opyene and Gedamu, 1997). Cathepsin B-like *cpc* also exists as a single copy in *L. mexicana* (Bart *et al*., 1995) and in *L. major* (Sakanari *et al*., 1997). In *T. congolense* 13 genes coding for cathepsin B-like cysteine proteases have been reported, where they do not occur in a tandem repeat, but are located on different chromosomes (Palomares *et al*., 2008), while only 1 copy exists in *T. brucei* and 2 in *T. vivax.*.

The existence of an *Mlu*I restriction site in the intergenic region of cysteine proteases of *T. congolense* made it possible to make a subgenomic library of complete gene units, unlike in previous experiments where restriction endonucleases such as *Pst*I and *Bam*HI were used, that cut within the coding sequence (leaving part of the C-terminal coding sequence). The latter restriction endonucleases both cut within the gene and therefore, generated partial cysteine protease gene libraries. It was not evident whether cysteine proteases that occur outside of the main locus were cloned through this approach, as all the clones obtained are closely related to those obtained earlier. All the clones from the subgenomic library had more than 86% identity with CP1, CP2 and CP3.

During the construction and screening of the 2 kb subgenomic library of congopain-like cysteine proteases a number of challenges were faced. 1/ For successful cloning, a completely cut and dephosphorylated vector is required. Purification by ethanol-precipitation without size fractionation on an agarose gel was found to generate high background upon transformation, because the cut vector was mixed with small amount of uncut vector. Uncut vector, under covalently closed circular conformation, is preferentially taken up by competent cells. Purification of the cut and dephosphorylated pMluI was found effective by separating the cut and uncut vector on an agarose gel and then excising and purifying the cut vector band. 2/ The introduction of a *Mlu*I site into the MCS of the pGEM®-11Zf(+) cloning vector by site-directed mutagenesis failed. The pMluI plasmid was constructed by an alternative method, involving cloning of annealed primers with the *Mlu*I restricted site into the pGEM®-11Zf(+) MCS. 3/ A high quantity of *T. congolense* IL 3000 genomic DNA was required for digestion with *Mlu*I restriction endonuclease to give sufficient insert. Hence large quantities of parasites were required for DNA extraction. 4/ It was difficult to establish the right vector to insert ratios. 5/ Lastly, very few congopain-like cysteine proteases coding clones were obtained from the library (8 out of 204, or about 4 %), the majority of the clones coding for non cysteine proteases.

It was noted that out of the seven clones obtained from the subgenomic library of 2 kbp fragments, only one clone had a cysteine residue at the active site while the rest had a serine residue. This observation was also made in a related study conducted in our laboratory, whereby a congopain gene-units library was constructed from the same cosmid clones but using the *Mlu*I restriction site (Pillay et al., manuscript in preparation): most clones obtained were found to be of the CP3 type, with a serine instead of a cysteine in the active site. This is

in contrast with the results reported in the Chapters 3 and 4, where the majority of the clones obtained had the expected cysteine residue as opposed to a serine residue in the active site. The main difference between the *Mlu*I clones and the others, either partial (PCR fragment) or cut in the coding sequence (like in the case of the *Pst*I library), is that these are complete gene units. There is a possibility that clones carrying congopain genes with a cysteine residue at the active site have the active enzyme produced in the bacteria, where it can prove lethal. Clones with a serine in the active site are likely to generate an inactive enzyme, hence are preferentially taken up by bacteria as not lethal.

The essential amino acid residues in the active site of cysteine protease are cysteine, histidine and asparagine as they form the catalytic triad (Storer and Ménard, 1994). Occasionally the substitution of a cysteine residue by a serine residue has been noted, as well as the substitution of other amino acid residues forming the catalytic triad (Kiefer *et al.*, 1996; Gor *et al.*, 1998). In a group of cysteine proteases called SERA in *Plasmodium*, the majority of genes carry a serine repeat and a serine residue instead of cysteine residue at the active site (Bzik *et al.*, 1988; Kiefer *et al.*, 1996; Gor *et al.*, 1998; Hodder *et al.*, 2003). The variability of cysteine protease genes has also been reported in plants where a protease named SPE31 isolated from the seeds of the legume *Pachyrhizus erosus* was found to have a glycine instead of the cysteine in the active site (Zhang *et al.*, 2006). It showed 78% sequence identity with protein P34 from soy beans that also has a glycine instead of a cysteine residue, but has no activity. Two clones from the subgenomic library constructed here had the histidine residue of the catalytic triad substituted by a serine residue. In *P. vivax* a similar observation was reported in SERA where the histidine residue was substituted by leucine (Kiefer *et al.*, 1996) and by methionine in *P. vinckei* (Gor *et al.*, 1998). These substitutions are believed to affect the proteolytic activities of these cysteine proteases and thus it has been suggested that some of these cysteine proteases could be the product of pseudogenes. Expression of some of the variants will be necessary to establish their activity. Miller *et al.* (2002) found that SERA-4 and 5 which contain a serine residue, is highly expressed in *P. falciparum*. A gene knock-out experiment would be necessary to determine their role in host-parasite interactions. SERA-4, 5 and 6 (with a cysteine residue at the active site) in *P. falciparum* were found to be necessary for the parasite's erythrocytic cycle, because these genes could not be disrupted by RNAi (Miller *et al.*, 2002). Also in *P. falciparum* cysteine proteases falcipain-1 and -2 activities was silenced by RNAi (Malhotra *et al.*, 2002).

RT-PCR confirmed the presence of cysteine protease mRNA in bloodstream forms of *T. congolense* IL 3000 for all these different isoforms, including the ones having a serine at the active site, suggesting that the different isoforms of congopain-like cysteine protease genes are expressed. The significance of the expression of these mutants remains unclear. The expression of some of these cysteine protease variants is underway in our laboratory and it will be interesting to characterise these enzymes, and to determine their role in parasite-host interactions by RNA interference.

Although there was a marked substitution of essential amino acid residues in clones obtained from the subgenomic library, it is important to note that these clones showed a high degree of identity (86-100%), which suggests that an immunogen based on congopain can serve as a vaccine candidate in addressing the pathology of trypanosomosis.

CHAPTER 6
GENERAL DISCUSSION

Trypanosomosis is the most important constraint to livestock production in sub-Saharan Africa (Kristjanson *et al.*, 1999). Control of trypanosomosis has largely depended on vector management (insecticides, tsetse traps) and trypanocidal drugs (Schofield and Maudlin, 2001) because they are considered relatively effective measures. Due to prolonged use of trypanocides, resistance to one or more of the trypanocidal drugs has been observed in African trypanosome-affected countries (Geerts *et al.*, 2001). Development of a conventional vaccine has been hampered by the ability of the parasite to evade host immune responses by both antigen variation and immunosuppression (Donelson, 2003). Use of trypanotolerant cattle in the tsetse infested areas has been successful (d'Ieteren *et al.*, 1998), but these cattle are smaller than their trypanosusceptible counterparts meaning that they have low pulling (tractor) power. Their milk and meat production is also lower compared to that of trypanosusceptible breeds.

Cysteine proteases are proteolytic enzymes present in almost all organisms, ranging from viruses to vertebrates (Sajid and McKerrow, 2002). Cysteine proteases of the papain-superfamily belong to the C1 family within clan CA (Barrett and Rawlings, 2001). They are located within the lysosomes and synthesised as zymogen precursors that are activated by cleavage of the pro-domain to generate mature enzymes. Cysteine protease inhibitors have been cited as potential chemotherapeutic agents in parasitic diseases (Sajid and McKerrow, 2002). In *Schistosoma* species cysteine protease based inhibitor peptidomimetic vinyl sulfone (K11777) exhibited at least 88% success in the elimination of parasite eggs and decrease in worm load (Abdulla *et al.*, 2007, Caffrey, 2007). Rosenthal *et al.*, (2004) observed that cysteine protease inhibitors block parasite haemoglobin degradation which is necessary for *P. falciparum* erythrocytic trophozoites, a role performed by falcipain-2.

Pathogenic factors of trypanosomes are increasingly becoming drug and vaccine targets, such as trypanopain-Tc from *T. congolense* (Mbawa *et al.*, 1992), congopain from *T. congolense* (Authié *et al.*, 2001; Lalmanach *et al.*, 2002), trypanopain-Tb from *T. b. brucei* (Troeberg *et al.*, 1999; Caffrey *et al.*, 2000) and cruzain of *T. cruzi* (Caffrey *et al.*, 2000). In *T. congolense*, a chimeric reversible inhibitor Pcp27 derived from the YHNGA inhibitory peptide of the pro-region was found to inhibit congopain (Godat, *et al.*, 2005). Tetramer

pseudopeptide reversible inhibitors were found to inhibit both cruzain and congopain (Lecaille *et al*., 2001). In *T. cruzi* and *T. brucei,* peptidyl and peptidomimetic inhibitors of brucipain and cruzipain have been found to limit parasite survival both *in vivo* and *in vitro* (Caffrey *et al*., 2000; Steverding *et al*., 2006).

Congopain, the major cysteine protease in *T. congolense* has been shown to contribute to the pathology of trypanosomosis (Authié *et al*., 1993, 2001; Authié, 1994). Cattle immunized with congopain develop high levels of anti-congopain IgG which inhibit its activity (Authié *et al*., 2001). Upon infection, immunised animals remained more productive than controls. Three cysteine proteinases of the papain family have been described in *T. congolense*, namely, CP1 (Fish *et al*., 1995, EMBL accession number Z25813), CP2 (Jaye *et al*., 1994, EMBL accession number L25130) and CP3 (Downey and Donelson, 1999 EMBL accession no. AF139913). The proteases predicted by the gene sequence of CP1 and CP2 are identical in their pre and pro-peptides, differing from CP3 by only two amino acid residues in the signal peptide, while their C-terminal extensions show 86-95% identity, and their catalytic domains are 88-95% identical. However they differ in their N-terminal sequences APPA for CP1 and CP3 and APEA for CP2 (Boulangé *et al*., 2001). CP3 differs from CP1 and CP2 due to a replacement of a cysteine residue at position Cys^{25} (papain numbering) by a serine residue. Development of an anti-disease vaccine or a chemotherapeutic agent based on congopain-like cysteine protease requires an understanding of the diversity of the genes coding for these proteases. Therefore the variability within these families of genes was studied firstly by constructing and screening a congopain-like-cysteine protease gene units library from some existing cosmid clones, secondly by amplification by PCR and cloning of cysteine proteases genes from *T. congolense*, *T. vivax* and *T. b. brucei* genomic DNA. Thirdly, by constructing and screening a subgenomic library. Finally the pattern of expression of these isoforms was investigated by RT-PCR.

Variability of congopain-like cysteine proteases was studied by constructing and screening a congopain-like-cysteine protease gene unit library from pre-existing congopain locus cosmid clones. The experiment was successful and 120 clones of congopain-like cysteine protease were obtained. It was shown that more than half of the clones were close to that of CP2, approximately a third to CP1 and the rest to CP3. The major amino acid sequence differences seem to occur in the catalytic domain and the proenzyme cleavage site sequences, APEA and APDA, with the former being predominant. Most of the clones coding for the APEA

sequence had a cysteine residue at the active site, while the clone coding for APDA had a serine instead. The other amino acid residues of the catalytic triad were conserved in all the clones. The C-terminal extension of congopain is joined to the catalytic domain by polyproline (Fish *et al.*, 1995; Boulangé *et al.*, 2001) as was the case in all the clones obtained showing six proline residues.

The second objective of the present study was to study the variability of cysteine proteases across strain and species. This was done through amplification of CP gene fragments from *T. congolense*, *T. vivax* and *T. brucei* genomic DNA, cloning into pGEM™-T Easy and analyses of the sequences to establish intra- and inter-species variability within congopain-like cysteine protease genes. A total of 87 genomic DNA samples from strains of the three species were obtained from various sources. Although degenerate primers were used initially for amplification, species-specific primers were found to be the most ideal for PCR. Across species, it was established that the intergenic region of *T. b. brucei* and *T. vivax* was shorter than that of *T. congolense*; in addition, *T. vivax* boasts two different sized intergenic regions, advocating for the presence of two loci for congopain-like genes ("vivapain"), a fact already reported elsewhere (Prof. T. Baltz, personal communication). The two bands were sequenced and found indeed to be part of cysteine protease genes. Amplification and cloning of part of the coding region was done successfully for *T. congolense* and *T. brucei spp.* However for *T. vivax* PCR reactions gave faint bands, and thus necessitated reamplification. The PCR products obtained after reamplification were not *T. vivax* cysteine proteases but those of *T. congolense* probably due to contamination from the original DNA. Therefore no coding sequence of *T. vivax* cysteine protease genes was obtained.

The region amplified from *T. congolense* and *T. b. brucei* coding regions was constituted by a part of the propeptide and the catalytic region. The proenzyme cleavage site varied from APEA to APPA, APDA or PPMT. The most common N-terminal sequence was APEA and found mostly in CPs with a cysteine residue at the active site. The PPMT sequence was rare and only found in some CPs with a serine instead of cysteine at the active site. The proportion of CP genes coding for a cysteine residue at the active site was higher than those coding for enzyme with a serine residue; one clone was found to have a phenylalanine in place of a cysteine. Cys/Ser mutants were not found in *T. brucei.* The C-terminal regions of *T. congolense* strains were linked to the catalytic domain by six proline residues; one clone had a unique hinge consisting of 14 proline residues. The peptide sequence in the propeptide

of cysteine protease responsible for inhibition of cysteine proteases (Lalmanach *et al.*, 1998) was conserved in all clones, being YHNGA for *T. congolense* and YRNGA for *T. b. brucei.*

The variability of congopain-like cysteine protease genes was further studied through construction and screening of a 2 kbp size-restricted subgenomic library from *T. congolense* IL 3000 genomic DNA. The subgenomic library was successfully constructed and yielded a total of 206 clones, amongst which eight were identified as coding for congopain-like cysteine proteases. Six clones were found coding for a serine residue at the active site instead of a cysteine and one clone coding for a cysteine which fell in the same cluster as CP1. The variability among the clones at the maturation site had a similar pattern as described above, with the APEA N-terminal sequence being dominant, APPA and APDA sequences were found in one clone each while the PPMT sequence was found in two clones. The asparagine residue of the catalytic triad was conserved in all the clones and substitution of histidine by serine was noted in two clones. Size polymorphism of the C-terminal extension was noticed among the clones, and difference in the number of amino acid residues forming the polyproline hinge emerged. Minor amino acid sequence differences among the clones were observed, Gly-19 (papain numbering) was substituted by arginine in three clones while it was conserved in the rest.

Although there are many restriction sites in the genes coding for cysteine proteases of *T. congolense*, not all have been used in our research, because of either their respective positions in the gene or their availability in the market, or both. Restriction sites in DNA sequences are frequently exploited in molecular biology, for genomic organisation studies, for library construction, or for polymorphism analyses. The restriction sites *Pst*I, *Sal*I, *Bam*HI, *Pvu*I and *Mlu*I that we used in cysteine protease genes have been used widely in genomic organisation studies in other organisms. These include *T. cruzi* (Eakin *et al.*, 1992), *L. major* (Sakanari *et al.*, 1997), *L. donovani chagasi* (Omara-Opyene and Gedamu, 1997), *L. donovani* complex (Mundodi *et al.*, 2002) and also in western flowers thrips cysteine protease (Kuipers and Jongsma, 2004). The genomic organisation of congopain was studied using *Pst*I, *Bam*HI and *Mlu*I restriction sites in the present study, the first two occur within the gene while the latter exist in the intergenic region of congopain. *Xba*I and *Hin*dIII sites were found in one clone each, previously both restriction were not thought to exist within the congopain-like cysteine protease genes (Dr. Boulangé personal communication), although this observation was based on sequence analysis of a very small number of congopain-like

cysteine protease genes. Both *Sal*I and *Nco*I sites were also found in the same clone; initially it was thought that *Sal*I was a marker of CP1-like CPs while *Nco*I was believed to be a marker of CP2 and related genes (Dr. Boulangé personal communication). The identification of these restriction sites might be useful in further studies on the cysteine protease isoforms of *T. congolense*.

Following the recurring identification of CP3-like sequences (where the active site Cys is replaced by Ser) in the present study, it was deemed necessary to determine whether CP3 – type of mutants were expressed in *T. congolense* parasites. We achieve that through RT-PCR. The results confirmed the presence of all types of cysteine protease mRNA in bloodstream forms of *T. congolense* IL 3000, suggesting that different isoforms of congopain-like cysteine proteases are indeed expressed. In *P. falciparum* microarray analysis, antigenicity studies and gene knockout were used in addition to RT-PCR to detect the expression of the nine SERA genes (Miller *et al.*, 2002).

It was shown in the present study that the variability within congopain-like cysteine protease genes occurs mainly in the sequences coding for the catalytic domain, while those coding for the signal peptide and the propeptide are strictly conserved. The translated C-terminal extension and the intergenic region sequences showed higher sequence identity than that of the catalytic domain. Variability was found at the proenzyme cleavage site coding sequences translating into several N-terminal sequences, i.e. APPA, APEA, APDA and PPMT in *T. congolense*, and only APAA in *T. b. brucei*. APEA was the most common N-terminal sequence found, while PPMT coding sequence was only found in a group of cysteine protease genes that also code for a serine residue at the active site. Variability at the active site was more common; more clones were obtained coding for a cysteine residue than those coding for serine and one clone was found that codes for phenylalanine. However, the different approaches used in the present study gave different proportions of genes coding for cysteine and serine at the active site. It was observed that more congopain-like cysteine protease genes coding for a serine than a cysteine residue were obtained from the size-restricted subgenomic library; contrary to results from the cosmid library and cloning of PCR products where congopain-like cysteine protease genes coding for cysteine were in the majority. In another study (outside the scope of the present study) where the cosmid clones (mentioned in Chapter five) were digested with *Mlu*I to generate a library, there was only one clone with a cysteine residue at the active site as opposed to many with serine residues.

These observations led to the hypothesis that when the entire gene unit encompassing regulatory sequences and open reading frame is cloned, some active enzyme is produced in the bacteria that have a deleterious effect on growth. This is thought not to happen when the gene is incomplete, as is the case in our other cloning (*Pst*I library, Chapter 3, or PCR fragments, Chapter 4), or when the enzyme is inactive, such as when the cysteine or histidine is substituted with an other amino acid.

Cysteine, histidine and asparagine, are the amino acid residues forming the catalytic triad of papain-like cysteine proteases (Storer and Ménard, 1994). In the present study, there was a marked cysteine to serine substitution at the active site of the catalytic domain, as well as substitution of other essential amino acid residues of the catalytic triad. The substitution of these essential amino acid residues was observed in congopain-like cysteine protease sequences obtained from http://www.genedb.org/genedb/tcongolense/blast.jsp. Other than the substitution of cysteine by a serine residue, histidine was substituted by either serine or tyrosine residues. Similarly, cysteine to serine substitutions were observed in cathepsin B-like cysteine proteases in *T. congolense* (Palomares *et al*., 2008); in this case six genes coded for cysteine while seven genes coded for a serine residue. The substitution of serine for cysteine has been observed in other protozoans. A group of papain-family cysteine proteases in *Plasmodium* called the serine-repeat antigens (SERA), has a serine residue instead of a cysteine at the active site (Bzik *et al*., 1988; Knapp *et al*., 1991; Kiefer *et al*., 1996; Gor *et al*., 1998; Hodder *et al*., 2003). Additionally, the substitution of other amino acid residues of the catalytic domain was also observed, in *P. vivax* where the histidine residue was substituted by leucine (Kiefer *et al*., 1996) and by methionine in *P. vinckei* (Gor *et al*., 1998). The variability of cysteine protease genes has also been reported in other organisms such as *Trichobilharzia regenti* (Dvořák *et al*., 2005) and also in plants. A protein named SPE31 isolated from the seeds of the legume *Pachyrhizus erosus* was found to have a glycine instead of the cysteine in the active site (Zhang *et al*., 2006) and showed 78% sequence identity with protein P34 from soy beans. These substitutions are believed to affect the proteolytic activities of these cysteine proteases. The substitution of very few amino acid residues in the catalytic domain of two cysteine protease isoforms (cruzain and cruzipain 2) from *T. cruzi* led to enzymes with different pH stability, substrate sensitivity and sensitivity to inhibition by cysteine protease inhibitors (Lima et al., 2001).

Expression of cysteine protease genes harbouring a serine residue at the catalytic site would be necessary to establish whether they are active. Likewise, gene knock-out studies would be necessary to determine the role of these groups of cysteine proteases in the pathogenesis of trypanosomosis. Miller *et al.* (2002) found that SERA-4 and 5 which contain a serine residue were highly expressed in *P. falciparum*, a similar observation was made for SERA-5 which was found predominantly expressed in the trophozoite and schizont stages (Aoki *et al.*, 2002). SERA-4, 5 and 6 (with a cysteine residue at the active site) of *P. falciparum* were found to be necessary for the erythrocytic cycle, because they could not be disrupted by RNAi (Miller *et al.*, 2002). RNA interference was successful in targeting the papain-like falcipain-1 or falcipain-2 activity in *P. falciparum*, thereby inhibiting the development of erythrocytic parasites and a marked accumulation of undegraded haemoglobin was observed (Malhotra *et al.*, 2002). SERA-5 was found to have serine protease (chymotrypsin-like) activity (Hodder *et al.*, 2003). Although SERA-6 has a conserved catalytic cysteine it does not have cysteine protease activity. However, due to localisation of the two SERA proteins to the parasitophorous vacuole of mature schizonts, they might be involved in proteolytic activities together with other SERA proteins (Rosenthal, 2004). Also, it has been suggested that some of these cysteine protease could be the products of pseudogenes.

The existence of serine to cysteine substitution at the active site of cysteine proteases has led to the suggestion that the enzymes might be encoded by pseudogenes (Pils and Schultz , 2004). In *T. congolense* two cathepsin B-like cysteine proteases were identified as being coded by pseudogenes (Palomares *et al.*, 2008). *Pyrococcus furiosus* 1-like genes coding for a cysteine peptidase were found to be expressed in *L. major* but not in other *Leishmania* species and was also absent in *T. brucei* and *T. cruzi* suggesting that they might exist as pseudogenes in other *Leishmania* species (Eschenlauer *et al.*, 2006). Calpains are a group of Ca^{2+}-dependent cytosolic cysteine proteases, which belong to the family C2 within the clan C2 of cysteine proteases. Substitution of catalytic cysteine to serine residue of the maize DEKI protein (calpain) made it inactive (Wang *et al.*, 2003), the same observation was made for other human calpain-like cysteine proteases (Masumoto *et al.*, 1998; Hitomi *et al.*, 1998). However, not all amino acid residue substitutions would render proteases inactive or be the products of pseudogene expression, especially residues outside the catalytic triad.

The catalytic domains coded by the congopain-like cysteine protease genes showed only few amino acid substitutions; some of the substitutions however are likely to affect the catalytic

characteristics of the enzyme, or its stability. Negatively charged amino acid residues (glutamate) were substituted by positively charged amino acid residues (arginine) or the amide asparagine; the hydrophilic amino acid residue aspartate was occasionally substituted by a hydrophobic phenylalanine. The differences in these genes may translate into different enzymes with different roles within the parasite, or in host-parasite interactions. In *T. congolense,* half of the amino acid substitutions encountered between C1 (CP1) and C2 (CP2) are a change from a basic residue towards an acid one, leading to the enzymes having very different p*I* values and elution profiles on ion-exchange chromatography (Boulangé *et al.*, 2001). Moreover CP2 exhibited activity over a wide pH range while CP1 was active only under acidic conditions, being inactivated at neutral or alkaline pH. Lima *et al.* (2001) reported that two cysteine protease isoforms (cruzain and cruzipain 2) from *T. cruzi* with very few amino acid substitutions in the catalytic domain led to enzymes with different pH stability, substrate sensitivity and sensitivity to inhibition by cysteine protease inhibitors. Indeed, cysteine proteases have a very specific substrates preference (Lecaille *et al.*, 1999, 2001; Juliano *et al.*, 2004, Judice *et al.*, 2005).

Cysteine proteases are expressed at high level in *L. mexicana* (McKerrow *et al.*, 2006). Three cysteine proteases of the papain family have been identified, namely CPA and CPC which are cathepsin L-like and CPB which is cathepsin B-like (Mottram *et al.*, 1998). The role of CPA in *L. mexicana* remains unclear as deletion of *cpa* genes had no effect in the development of lesions in mice as well as macrophage infection (Souza *et al.*, 1994). CPC on the other hand, plays a major role in macrophage infections; mutants for the *cpc* genes in *L. mexicana* showed impaired macrophage infection and affected lesion development *in vivo* (Bart *et al.*, 1997). The role of CPB in pathogenesis in *L. mexicana* is more profound than that of either CPA or CPC (McKerrow *et al.*, 2006). CPB has been shown to facilitate macrophage infections as CPB-1 and CPB-2 are highly expressed in infectious metacyclic promastigotes stages of *L. mexicana* (Mottram *et al.*, 1997). Lastly CPB has been shown to inhibit Th1 responses and also plays a key role in suppressing protective immune responses in *L. mexicana* (Buxbaum *et al.*, 2003).

Cysteine proteases are important virulence factors in *E. histolytica* as they are required for parasite survival during acute amoebic liver abscesses in hamsters (Olivos-Garcia *et al.*, 2004). A total of 20 cysteine proteases have been predicted to exist in *E. histolytica* (Bruchhaus *et al.*, 2003), but only three of these cysteine proteases have been found to have

activity, namely EhCP1, EhCP2 and EhCP5 (Bruchhaus *et al.*, 1996). EhCP2 and EhCP5 are important for enteric cytopathology because they contribute to host layer disruption and EhCP2 also facilitates pathologic host inflammatory responses (Hellberg *et al.*, 2001; Zhang *et al.*, 2000).

Falcipains are a group of cysteine proteases of the papain-family encoded by *P. falciparum* (Sijwali *et al.*, 2006). Falcipain-1 was isolated and characterised (Rosenthal and Nelson, 1992), and thought to be involved in host cell invasion (Greenbaum *et al.*, 2002), although gene disruption studies have shown that this protease is not necessary for erythrocytic stage parasites (Eksi *et al.*, 2004; Sijwali *et al.*, 2004, 2006). Falcipain-2 has been linked with degradation of haemoglobin (Shenai *et al.*, 2000; Sijwali *et al.*, 2000; Dahl and Rosenthal, 2005). Indeed gene knockout studies indicate the role of falcipain-2 in haemoglobin degradation, as parasites showed reduced haemoglobin hydrolysis in trophozoites (Sijwali and Rosenthal, 2004). Sijwali *et al.* (2006) observed that falcipain-1 and falcipain-2' do not play an important role in haemoglobin hydrolysis (the role of falcipain-2) and in erythrocytic parasites (the role of falcipain-3).

The understanding of the roles of congopain, congopain-like cysteine proteases, and other virulence factors in pathogenesis will be a big step towards development of new drugs or vaccines against African trypanosomosis. The present study was important as it highlighted the diversity of the congopain-like gene family. Sequence alignments led to the identification of amino acid residues that may play a major role in the substrate specificity of the enzymes, a fact that may prove crucial in the design of chemotherapeutic inhibitors. Although there is a degree of variability in congopain-like cysteine proteases, predicted amino acid sequences are 88-100% identical. It is highly likely that a congopain-based vaccine will inhibit all congopain-like cysteine proteases that may be circulating in the infected host. This study highlighted also the ubiquitous substitution of a cysteine by a serine in the active site of several of these gene products. Further studies on the activity of such enzymes, and their potential role in host-parasite interactions will have to be tackled using recombinant technologies and reverse genetics.

REFERENCES

Abdulla, M.H., **Lim, K.C.**, **Sajid, M.**, **McKerrow, J.H. and Caffrey, C.R**. (2007). *Schistosomiasis mansoni*: novel chemotherapy using a cysteine protease inhibitor. *PLoS. Med*. **4,** e14.

Alvarez, F., Cortinas, M.N. and Musto, H. (1996). The analysis of protein coding genes suggests monophyly of *Trypanosoma*. *Mol. Phylogenet. Evol*. **5,** 333-343.

Aoki, S., Itagaki, S., Okech, B.A., Egwang, T.G., Matsuoka, H., Palacpac, N.M., Mitamura, T. and Horii, T. (2002). Serine repeat antigen (SERA5) is predominantly expressed among the SERA multigene family of *Plasmodium falciparum*, and the acquired antibody titers correlate with serum inhibition of the parasite growth. *J. Biol. Chem*. **277,** 47533-47540.

Asbeck, K., Ruepp, S., Roditi, I. and Gibson, W. (2000). GARP is highly conserved among *Trypanosoma congolense* Savannah, Forest and Kilifi subgroups. *Mol. Biochem. Parasitol*. **106,** 303-306.

Authié, E. (1994). Trypanosomosis and trypanotolerance in cattle: a role for congopain? *Parasitol. Today*. **10**, 360-364.

Authié, E. and Pobel, L. (1990). Serum haemolytic complement activity and C3 levels in bovine trypanosomosis under natural conditions of challenge-early indications of individual susceptibility to disease. *Vet. Parasitol*. **35,** 43-59.

Authié, E., Muteti, D.K., Mbawa, Z.R., Lonsdale-Eccles, J.D., Webster, P. and Well, C.W. (1992). Identification of a 33-kilodalton immunodominant antigen of *Trypanosoma congolense* as a cysteine protease. *Mol. Biochem. Parasitol*. **56**, 103-116.

Authié, E., Duvallet, G., Robertson, C. and **Williams, D.J.** (1993). Antibody response to a 33kDa cysteine protease of *Trypanosoma congolense*: relationship to ''trypanotolernce'' in cattle. *Parasite immunol*. **15**, 465-474.

Authié, E., Boulangé, A., Muteti, D., Lalmanach, G., Gauthier, F. and Musoke, A. J. (2001). Immunisation of cattle with cysteine proteinases of *Trypanosoma congolense*: Targeting the disease rather than the parasite. *Int. J. Parasitol*. **31**, 1429-1433.

Barrett, A.J. (1994). Classification of peptidases. *Methods Enzymol*. **244,** 1-15.

Barrett, A.J. and Rawlings, N.D. (2001). Evolutionary lines of cysteine peptidases. *Biol. Chem*. **382,** 727-733.

Bart, G., **Coombs, GH. and Mottram, J.C**. (1995). Isolation of *lmcpc*, a gene encoding a *Leishmania mexicana* cathepsin-B-like cysteine proteinase. *Mol. Biochem. Parasitol*. **73,** 271-274.

Bart, G., **Frame, M.J.**, **Carter, R.**, **Coombs, G.H. and Mottram, J.C.** (1997). Cathepsin B-like cysteine proteinase-deficient mutants of *Leishmania mexicana*. *Mol. Biochem. Parasitol*. **88,** 53-61.

Berti, P.J. and Storer, A.C. (1995). Alignment/phylogeny of the papain superfamily of cysteine proteases. *J. Mol. Biol*. **246,** 273-283.

Birnboim, H.C. and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res*. **7**, 1513-1523.

Blackman, M.J. and Holder, A.A. (1992). Secondary processing of the *Plasmodium falciparum* merozoite surface protein-1 (MSP1) by a calcium-dependent membrane-bound serine protease: shedding of MSP1$_{33}$ as a noncovalently associated complex with other fragments of the MSP1. *Mol. Biochem. Parasitol*. **50,** 307-315.

Boom, R., Sol, C.J., Salimans, M.M., Jansen, C.L., Wertheim-van Dillen, P.M. and van der Noordaa, J. (1990). Rapid and simple methods for purification of nucleic acids. *J. Clinical Microbiol*. **28,** 495-503.

Boulangé, A., Serveau, C., Brillard, M., Minet, C., Gauthier, F., Diallo, A., Lalmanach, G. and Authié, E. (2001). Functional expression of the catalytic domains of two cysteine proteinases from *Trypanosoma congolense*. *Int. J. Parasitol*. **31**, 1435-1440.

Bruchhaus, I., **Jacobs, T.**, **Leippe, M**. **and Tannich, E**. (1996). *Entamoeba histolytica* and *Entamoeba dispar*: differences in numbers and expression of cysteine proteinase genes. *Mol. Microbiol*. **22,** 255-263.

Bruchhaus, I. and Tannich, E. (1996). A gene highly homologous to ACP1 encoding cysteine proteinase 3 in *Entamoeba histolytica* is present in and expressed in *E.dispar*. *Parasitol. Res*. **82**, 189-192.

Bruchhaus, I., **Loftus, B.J.**, **Hall, N. and Tannich, E**. (2003). The intestinal protozoan parasite *Entamoeba histolytica* contains 20 cysteine protease genes, of which only a small subset is expressed during *in vitro* cultivation. *Eukaryot. Cell.* **2,** 501-509.

Buxbaum, L.U., **Denise, H.**, **Coombs, G.H.**, **Alexander, J.**, **Mottram, J.C. and Scott, P**. (2003). Cysteine protease B of *Leishmania mexicana* inhibits host Th1 responses and protective immunity. *J. Immunol.* **171,** 3711-3717.

Bzik, D.J., Li, W.B., Horii, T. and Inselburg, J. (1988). Amino acid sequence of the serine-repeat antigen (SERA) of *Plasmodium falciparum* determined from cloned cDNA. *Mol. Biochem. Parasitol.* **30,** 279-288.

Caffrey, C.R. (2007). Chemotherapy of schistosomiasis: present and future. *Curr. Opin. Chem. Biol.* **11,** 433-439. .

Caffrey, C.R., Scory, S. and Steverding, D. (2000). Cysteine proteinases of trypanosome parasites: novel targets for chemotherapy. *Curr. Drug. Targets.* **1,** 155-162.

Cazzulo, J.J., Couso, R., Raimondi, A., Wernestedt, C. and Hellman, U. (1989). Further characterisation and partial amino acid sequence of a cysteine proteinase from *Trypanosoma cruzi. Mol. Biochem. Parasitol.* **33**, 33-41.

Cazzulo, J.J., Cazzulo Franke, M.C., Martinez, J. and Franke de Cazzulo, B.M. (1990). Some kinetics properties of a cysteine proteinase (cruzipain) from *Trypanosoma cruzi. Biochim. Biophys. Acta*. **1037**, 186-191.

Chagas, J.R., Authié, E., Serveau, C., Lalmanach, G., Juliano, L. and Gauthier, F. (1997). A comparison of the enzymatic properties of the major cysteine proteinases from *Trypanosoma congolense* and *Trypanosoma cruzi. Mol. Biochem. Parasitol*. **88,** 85-94.

Chen, D.Q., **Kolli, B.K.**, **Yadava, N.**, **Lu, H.G.**, **Gilman-Sachs, A.**, **Peterson, D.A. and Chang, K.P**. (2000). Episomal expression of specific sense and antisense mRNAs in *Leishmania amazonensis*: modulation of gp63 level in promastigotes and their infection of macrophages *in vitro. Infect. Immun.* 68, 80-86.

Clark, J.M. (1988). Novel non-template nucleotide addition reactions catalysed by Prokaryotic and Eukaryotic DNA polymerase. *Nucleic Acids Res.* **16**, 9677- 9686.

Coetzer, T.H., **Goldring, J.P. and Huson, L.E.** (2008). Oligopeptidase B: A processing peptidase involved in pathogenesis. <u>*Biochimie.*</u> **90,** 336-344.

Dagert, M. and Ehrlich, S.D. (1979). Prolonged incubation in calcium chloride improves the competence of *Escherichia. coli* cells. *Gene*. **6,** 23-28.

Dahl, E.L. and Rosenthal, P.J. (2005). Biosynthesis, localization, and processing of falcipain cysteine proteases of *Plasmodium falciparum. Mol. Biochem. Parasitol.* **139,** 205-212.

Dale, J.W. and von Schantz, M. (2002). From Genes to Genomes: Concepts and Applications of DNA Technology. *1ˢᵗ Edition* pp 99-116.

De Souza Leao, S., Lang, T., Prina, E., Hellio, R. and Antoine, J.C. (1995). Intracellular *Leishmania amazonensis* amastigotes internalize and degrade MHC class II molecules of their host cells. *J. Cell Sci.* **108**, 3219-3231.

Desmazes, C., Gauthier, F. and Lalmanach, G. (2001). Cathepsin L, but not cathepsin B, is a potential kininogenase. *Biol. Chem*. **382,** 811-815.

Desquesnes, M. (1997). Evaluation of a simple PCR technique for the diagnosis of *Trypanosoma vivax* infection in the serum of cattle in comparison to parasitological techniques and antigen-enzyme-linked immuno sorbent assay. *Acta Trop*. **65,** 139-148.

d'Ieteren, G.D., Authié, E., Wissocq, N. and Murray, M. (1998). Trypanotolerance, an option for sustainable livestock production in areas at risk from trypanosomosis. *Rev. Sci. Tech*. **17,** 154-175.

Donelson, J.E. (2003). Antigenic variation and the African trypanosome genome. *Acta Trop*. **85,** 391-404.

Downey, N. and Donelson, J.E. (1999). Expression of foreign proteins in *Trypanosoma congolense. Mol. Biochem. Parasitol*. **104,** 39-53.

Eakin, A.E., Mills, A.A., Harth, G., McKerrow, J.H. and Craik, C.S. (1992). The sequence, organisation and expression of the major cysteine protease (Cruzian) from *Trypanosoma cruzi. J. Biol. Chem.* **267**, 7411-7420.

Eakin, A.E., McGrath, M.E., McKerrow, J.H., Fletterick, R.J. and Craik, C.S. (1993). Production of cystallizable cruzain, the major cysteine protease from *Trypanosoma cruzi. J. Biol. Chem.* **268**, 6115-6118.

Eggleson, K.K., Duffin, K.L. and Goldberg, D.E. (1999). Identification and characterization of falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite *Plasmodium falciparum. J. Biol. Chem.* **274**, 32411-32417.

Eksi, S., **Czesny, B.**, **Greenbaum, D.C.**, **Bogyo, M. and Williamson, K.C**. (2004). Targeted disruption of *Plasmodium falciparum* cysteine protease, falcipain 1, reduces oocyst production, not erythrocytic stage growth. *Mol. Microbiol.* **53,** 243-250.

El-Sayed, N.M., Hedge, P., Quackenbush, J., Melville, S.E. and Donelson, J.E. (2000). The African trypanosome genome. *Int. J. Parasitol.* **30,** 329-345.

Eschenlauer, S.C., Coombs, G.H. and Mottram, J.C. (2006). PFPI-like genes are expressed in *Leishmania major* but are pseudogenes in other *Leishmania* species. *FEMS Microbiol. Lett*. **260,** 47-54.

Fish, W.R., Nkhungulu, Z.M., Muriuki, C.W., Ndegwa, D. M., Lonsdale-Eccles, J.D. and Steyaert, J. (1995). Primary structure and partial characterization of a life-cycle-regulated cysteine protease from *Trypanosoma (Nannomonas) congolense*. *Gene*. **161**, 125-128.

Frommel, T.O., **Button, L.L.**, **Fujikura, Y. and McMaster, W.R**. **(1990).** The major surface glycoprotein (GP63) is present in both life stages of *Leishmania. Mol. Biochem. Parasitol.* **38,** 25-32.

Garcia-Rivera, G., Rodriguez, M. A., Ocádiz, R., Martinez-López, M. C., Arroyo, R., González-Robles, A. and Orozco, E. (1999). *Entamoeba histolytica*: a novel cysteine protease and an adhesin form the 112 kDa surface protein. *Mol. Microbiol.* **33**, 556-568.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D,J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M. and Barrell B. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum. Nature* **419,** 498-511.

Gashumba, J.K., Baker, R.D. and Goldfrey, D.G. (1988). *Trypanosoma congolense*: the distribution of enzymic variants in east and west Africa. *Parasitol.* **96**, 475-486.

Geerts, S., Holmes, P.H., Eisler, M.C. **and Diall, O.** (2001). African bovine trypanosomosis: the problem of drug resistance. *Trends Parasitol.* **17**, 25-28.

Gibson, W.C., Stevens, J.R., Mwendia, C.M., Ngotho, J.N. and Ndung'u, J.M. (2001). Unravelling the phylogenetic relationships of African trypanosomes of suids. *Parasitol.* **122,** 625-631.

Godat, E., **Chowdhury, S.**, **Lecaille, F.**, **Belghazi, M.**, **Purisima, E.O. and Lalmanach, G**. (2005). Inhibition of a cathepsin L-like cysteine protease by a chimeric propeptide-derived inhibitor. *Biochemistry*. **44,** 10486-10493.

Goldberg, D.E., Slater, A.F., Beavis, R., Chait, B.M., Cerami, A. and Henderson, G.B. (1991). Hemoglobin degradation in the human malaria pathogen *Plasmodium falciparum*: a catabolic pathway initiated by a specific aspartic protease. *J. Exp. Med.* **173**, 961-969.

Gómez, E., **Valdés, A.M.**, **Piñero, D. and Hernández, R**. (1991). What is a genus in the Trypanosomatidae family? Phylogenetic analysis of two small rRNA sequences. *Mol. Biol. Evol*. **8,** 254-259.

Gor, D.O., Li, A.C. and Rosenthal, P.J. (1998). Protective immune responses against protease-like antigens of the murine malaria parasite *Plasmodium vinckei. Vaccine*. **16,** 1193-2202.

Greenbaum, D.C., **Baruch, A.**, **Grainger, M.**, **Bozdech, Z.**, **Medzihradszky, K.F.**, **Engel, J.**, **DeRisi, J.**, **Holder, A.A. and Bogyo, M**. (2002). A role for the protease falcipain 1 in host cell invasion by the human malaria parasite. *Science.* **298,** 2002-2006.

Haag, J., O'hUigin, C. and Overath, P. (1998). The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria. *Mol. Biochem. Parasitol*. **91,** 37-49.

Hannaert, V., Blaauw, M., Kohl, L., Allert, S., Opperdoes, F.R. and Michels, P.A. (1992). Molecular analysis of the cytosolic and glycosomal glyceraldehyde-3-phosphate dehydrogenase in *Leishmania mexicana. Mol. Biochem. Parasitol*. **55,** 115-126.

Hannaert, V., Opperdoes, F.R. and Michels, P.A. (1998). Comparison and evolutionary analysis of the glycosomal glyceraldehyde-3-phosphate dehydrogenase from different Kinetoplastida. *J. Mol. Evol*. **47,** 728-738.

Hellberg, A., **Nickel, R.**, **Lotter, H.**, **Tannich, E**. **and Bruchhaus I**. (2001). Overexpression of cysteine proteinase 2 in *Entamoeba histolytica* or *Entamoeba dispar* increases amoeba-induced monolayer destruction *in vitro* but does not augment amoebic liver abscess formation in gerbils. *Cell. Microbiol.* **3,** 13-20.

Hide, G., Welburn, S.C., Tait, A. and Maudlin, I. (1994). Epidemiological relationship of *Trypanosoma brucei* stocks from South-east Uganda: evidence for different population structures in human infective non-human infective isolates. *Parasitol.* **109**, 95-111.

Hitomi, K., **Uchiyama, Y.**, **Ohkubo, I.**, **Kunimatsu, M.**, **Sasaki, M. and Maki, M.** (1998). Purification and characterization of the active-site-mutated recombinant human mu-calpain expressed in baculovirus-infected insect cells. *Biochem. Biophys. Res. Commun.* **246,** 681-685.

Hodder, A.N., Drew, D.R., Epa, V.C., Delorenzi, M., Bourgon, R., Miller, S.K., Moritz, R.L., Frecklington, D.F., Simpson, R.J., Speed, T.P., Pike, R.N. and Crabb, B.S. (2003). Enzymic, phylogenetic, and structural characterization of the unusual papain-like protease domain of *Plasmodium falciparum* SERA5. *J. Biol. Chem*. **278,** 48169-48177.

Hu, G. (1993). DNA polymerase-catalysed addition of nontemplated extra nucleotides to the 3'end of a DNA fragment. *DNA Cell Biol.* **12**, 763-770.

Huete-Pérez, J.A., Engel, J.C., Brinen, L.S., Mottram, J.C. and McKerrow, J.H. (1999). Protease trafficking in two primitive eukaryotes is mediated by a prodomain protein motif. *J. Biol. Chem*. **274**, 16249-16256.

Inoue, H., Nojima, H. and Okayama, H. (1990). High efficiency transformation of *Escherichia coli* with plasmids. *Gene*. **96**, 23-28.

Israel, D.I. (1993). A PCR-based method for high stringency screening of DNA libraries. *Nucleic Acids Res*. **21,** 2627-2631.

Joshi, P.B., **Kelly, B.L.**, **Kamhawi, S.**, **Sacks, D.L. and McMaster, W.R**. (2002). Targeted gene deletion in *Leishmania major* identifies leishmanolysin (GP63) as a virulence factor. *Mol. Biochem. Parasitol.* **120,** 33-40.

Judice, W.A., **Mottram, J.C.**, **Coombs, G.H.**, **Juliano, M.A. and Juliano. L**. (2005). Specific negative charges in cysteine protease isoforms of *Leishmania mexicana* are highly influential on the substrate binding and hydrolysis. *Mol. Biochem. Parasitol.* **144,** 36-43.

Juliano, M.A., Brooks, D.R., Selzer, P.M., Pandolfo, H.L., Judice, W.A., Juliano, L., Meldal, M., Sanderson, S.J., Mottram, J.C. and Coombs. G.H. (2004). Differences in substrate specificities between cysteine protease CPB isoforms of *Leishmania mexicana* are mediated by a few amino acid changes. *Eur. J. Biochem.* **271,** 3704-3714.

Kiefer, M.C., Crawford, K.A., Boley, L.J., Landsberg, K.E., Gibson, H.L., Kaslow, D.C. and Barr, P.J. (1996). Identification and cloning of a locus of serine repeat antigen (sera)-related genes from *Plasmodium vivax. Mol. Biochem. Parasitol.* **78,** 55-65.

Klemba, M. and Goldberg, D. E. (2002). Biological roles of proteases in parasitic protozoa *Annu. Rev. Biochem*. **71**, 275-305.

Knapp, B., Nau, U., Hundt, E. and Küpper, H.A. (1991). A new blood stage antigen of *Plasmodium falciparum* highly homologous to the serine-stretch protein SERP. *Mol. Biochem. Parasitol.* **44,** 1-13.

Kristjanson, P.M., Swallow, B.M., Rowlands, G.J., Kruska, R.L. and **de Leeuw, P.N.** (1999). Measuring the costs of African animal trypanosomosis: the potential benefits of control and returns to research. *Agric. Syst*. **59,** 79-98.

Kuipers, A.G. and **Jongsma, M.A.** (2004). Isolation and molecular characterisation of cathepsin L-like cysteine protease cDNAs from western flower thrips (*Frankliniella occidentalis*). *Comp. Biochem. Physiol*. **139**, 65-75.

Kuru, T., **Jirata, D.**, **Genetu, A.**, **Barr, S.**, **Mengistu, Y.**, **Aseffa, A. and Gedamu, L**. (2007). *Leishmania aethiopica*: identification and characterization of cathepsin L-like cysteine protease genes. *Exp. Parasitol.* **115,** 283-290.

Lalmanach, G., **Lecaille, F.**, **Chagas, J.R.**, **Authié, E.**, **Scharfstein, J.**, **Juliano, M.A. and Gauthier, F.** (1998). Inhibition of trypanosomal cysteine proteinases by their propeptides. *J. Biol. Biochem.* **273,** 25112-25116.

Lalmanach, G., **Boulangé, A.**, **Serveau, C.**, **Lecaille, F.**, **Scharfstein, J.**, **Gauthier, F. and Authié, E**. (2002). Congopain from *Trypanosoma congolense*: drug target and vaccine candidate. *Biol. Chem.* **383**, 739-749.

Lecaille, F., **Serveau, C.**, **Gauthier, F. and Lalmanach, G.** (1999). Revisiting the S2 specificity of papain by structural analogs of Phe. *FEBS Lett.* **445,** 311-314.

Lecaille, F., **Cotton, J.**, **McKerrow, J.H.**, **Ferrer-Di Martino, M.**, **Boll-Bataillé, E.**, **Gauthier., F. and Lalmanach, G**. (2001). Reversible inhibition of cathepsin L-like proteases by 4-mer pseudopeptides. *FEBS Lett.* **507,** 362-366.

Lima, A.P., **dos Reis, F.C.**, **Serveau, C.**, **Lalmanach, G.**, **Juliano, L.**, **Ménard, R.**, **Vernet, T.**, **Thomas, D.Y.**, **Storer, A.C. and Scharfstein, J**. (2001). Cysteine protease isoforms from *Trypanosoma cruzi*, cruzipain 2 and cruzain, present different substrate preference and susceptibility to inhibitors. *Mol. Biochem. Parasitol.* **114,** 41-52.

Lonsdale-Eccles, J.D. and Mpimbaza, G.W. (1986). Thiol- dependent proteases of African trypanosomes. Analysis by electrophoresis in sodium dodecyl sulphate/polyacrylamide gels co-polymerised with fibrinogen. *Eur. J. Biochem.* **155**, 469-473.

Lonsdale-Eccles, J. D. and Grab, D.J. (1987). Lysosomal and non-lysosomal peptidyl hydrolases of the bloodstream forms of *Trypanosoma brucei brucei. Eur. J. Biochem.* **169**, 467-475.

Lukeš, J., **Jirků, M.**, **Dolezel, D.**, **Kral'ová, I.**, **Hollar, L. and Maslov, D.A**. (1997). Analysis of ribosomal RNA genes suggests that trypanosomes are monophyletic. *J. Mol. Evol.* **44,** 521-527.

Malhotra, P., **Dasaradhi, P.V.**, **Kumar, A.**, **Mohammed, A.**, **Agrawal, N.**, **Bhatnagar, R.K. and Chauhan, V.S**. (2002). Double-stranded RNA-mediated gene silencing of cysteine proteases (falcipain-1 and -2) of *Plasmodium falciparum. Mol. Microbiol.* **45,** 1245-1254.

Marko, M.A., **Chipperfield, R. and Birnboim, H.C**. (1982). A procedure for the large-scale isolation of highly purified plasmid DNA using alkaline extraction and binding to glass powder. *Anal. Biochem.* **121**, 382-387.

Masiga, D.K., **Smyth, A.J.**, **Hayes, P.**, **Bromidge, T.J. and Gibson, W.C**. (1992). Sensitive detection of trypanosomes in tsetse flies by DNA amplification. *Int. J. Parasitol.* **22,** 909-918.

Masiga, D.K., **McNamara, J.J. and Gibson, W.C.** (1996). A repetitive DNA sequence specific for *Trypanosoma (Nannomonas) godfreyi. Vet. Parasitol.* **62**, 27-33.

Masiga, D.K., **Tait, A. and Turner, C.M.** (2000). Amplified restriction fragment length polymorphism in parasite genetics. *Parasitol. Today*. **16**, 350-353.

Maslov, D.A., **Lukeš, J.**, **Jirků, M. and Simpson, L.** (1996). Phylogeny of trypanosomes as inferred from the small and large subunit rRNAs: implications for the evolution of parasitism in the trypanosomatid protozoa. *Mol. Biochem. Parasitol.* **75**, 197-205.

Masumoto, H., **Yoshizawa, T.**, **Sorimachi, H.**, **Nishino, T.**, **Ishiura, S. and Suzuki, K**. (1998). Overexpression, purification, and characterization of human m-calpain and its active site mutant, m-C105S-calpain, using a baculovirus expression system. *J. Biochem.* **124,** 957-961.

Matthews, K.R., **Tschudi, C. and Ullu, E.** (1994). A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes Dev.* **8,** 491-501.

Mattioli, R.C., **Pandey, V.S.**, **Murray, M. and Fitzpatrick, J.L.** (2000). Immunogenetic influences on tick resistance in African cattle with particular references to trypanotolerant N'Dama (*Bos Taurus*) and trypanosusceptible Gobra zebu (*Bos indicus*) cattle. *Acta trop.* **75**, 263-277.

Maxam, A.M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* **74**, 560-564.

Mayer, R., Picard, I., Lawton, P., Grellier. P., Barrault, C., Monsigny, M. and Schrével, J. (1991). Peptide derivatives specific for a *Plasmodium falciparum* proteinase inhibit the human erythrocyte invasion by merozoites. *J. Med. Chem.* **34**, 3029-3035.

Mbawa, Z.R., Gumm, I.D., Shaw, E. and Lonsdale-Eccles, J.D. (1992). Characterisation of a cysteine protease from bloodstream forms of *Trypanosoma congolense. Eur. J. Biochem.* **204,** 371-379.

McDermott, J.J and Coleman, P.G. (2001). Comparing apples and oranges-model-based assessment of different tsetse-transmitted trypanosomosis control strategies. *Int. J. Parasitol.* **31,** 603-609.

McGwire, B. and Chang, K. P. (1994). Genetic rescue of surface metalloproteinase (gp63)- deficiency in *Leishmania amazonensis* variants increases their infection of macrophages at the early phase. *Mol. Biochem. Parasitol.* **66**, 345-347.

McKerrow, J.H. (1999). Development of cysteine protease inhibitors as chemotherapy for parasitic diseases: insights on safety, target validation, and mechanism of action. *Int. J. Parasitol.* **29,** 833-837.

McKerrow, J.H., Engel, J.C. and Caffrey, C.R. (1999). Cysteine protease inhibitors as chemotherapy for parasitic infections. *Bioorg. Med. Chem.* **7,** 639-644.

McKerrow, J.H., **Caffrey, C.**, **Kelly, B.**, **Loke, P. and Sajid, M**. (2006). Proteases in parasitic diseases. *Annu. Rev. Pathol.* **1,** 497-536.

Melville, S.E., Leech, V., Gerrard, C.S., Tait, A. and Blackwell, J.M. (1998). The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* and assignment of the chromosomes makers. *Mol. Biochem. Parasitol.* **94***,* 153-173.

Melville, S.E., Leech, V., Navarro, M. and Cross, G.A. (2000). The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* stock 427. *Mol. Biochem. Parasitol.* **111**, 261-273.

Miller, S.K., Good, R.T., Drew, D.R., Delorenzi, M., Sanders, P.R., Hodder, A.N., Speed, T.P., Cowman, A.F., de Koning-Ward, T.F. and Crabb, B.S. (2002). A subset of *Plasmodium falciparum* SERA genes are expressed and appear to play an important role in the erythrocytic cycle. *J. Biol. Chem.* **277**, 47524-47532.

Morty, R.E., **Lonsdale-Eccles, J.D.**, **Morehead, J.**, **Caler, E.V.**, **Mentele, R.**, **Auerswald, E.A.**, **Coetzer, T.H.**, **Andrews, N.W. and Burleigh, B.A**. (1999). Oligopeptidase B from *Trypanosoma brucei*, a new member of an emerging subgroup of serine oligopeptidases. *J. Biol. Chem.* **274,** 26149-26156.

Morty, R.E., Pellé, R., Vadasz, I., Uzcanga, G.L., Seeger, W. and Bubis, J. (2005). Oligopeptidase B from *Trypanosoma evansi*. A parasite peptidase that inactivates atrial natriuretic factor in the bloodstream of infected hosts. *J. Biol. Chem.* **280,** 10925-10937.

Morty, R.E., Bulau, P., Pellé, R., Wilk, S. and Abe, K. (2006). Pyroglutamyl peptidase type I from *Trypanosoma brucei*: a new virulence factor from African trypanosomes that de-blocks regulatory peptides in the plasma of infected hosts. *Biochem. J.* **394,** 635-645.

Mottram, J.C., North, M.J., Barry, J.D. and Coombs, G.H. (1989). A cysteine protease cDNA from *Trypanosoma brucei* predicts an enzyme with an unusual C-terminal extension *FEBS Lett.* **258,** 211-215.

Mottram, J.C., **Robertson, C.D.**, **Coombs, G.H. and Barry, J.D**. (1992). A developmentally regulated cysteine proteinase gene of *Leishmania mexicana. Mol. Microbiol.* **6,** 1925-1932.

Mottram, J.C., **Frame, M.J.**, **Brooks, D.R.**, **Tetley, L.**, **Hutchison, J.E.**, **Souza, A.E. and Coombs, G.H**. (1997). The multiple *cpb* cysteine proteinase genes of *Leishmania mexicana* encode isoenzymes that differ in their stage regulation and substrate preferences. *J. Biol. Chem.* **272,** 14285-14293.

Mottram, J.C., Brooks, D.R. and Coombs, G.H. (1998). Roles of cysteine proteinases of trypanosomes and *Leishmania* in host-parasite interactions. *Curr. Opin. Microbiol.* **1**, 455-460.

Mullis, K.B. and **Faloona, F.A**. (1987). Specific synthesis of DNA *in vitro* via a polymerase-catalysed chain reaction. *Methods Enzymol.* **155**, 335-350.

Mundodi, V., Somanna, A., Farrel, P.J. and **Gedamu, L**. (2002). Genomic organisation and functional expression of differentially regulated cysteine protease genes of *Leishmania donovani* complex. *Gene.* **282,** 257-265.

Murray, M., Morrison, W.I. and Whitelaw, D.D. (1982). Host susceptibility of African trypanosomiasis: trypanotolerance. In: Baker, J. and Muller, R. (Eds) *Adv. Parasitol.* **21,** 1-68.

Murray, M. and **Dexter, T.M**. (1988). Anaemia in bovine African trypanosomosis. *Acta Trop.* **45**, 389-432.

Murray, M., Stear, M.J., Trail, J.C.M., dIetern, G.D.M., Agyemang, K. and **Dwinger, R.H.** (1991). Trypanosomosis in cattle: prospects for control. In: Axford, R. F. E. and Owen, J. B. (Eds) Breeding for disease resistance in Farm Animals. *CAB International, Wallingford, UK*, pp. 203-223.

Myers, T.W. and Gelfand, D.H. (1991). Reverse transcription and DNA amplification by a *Thermus thermophilus* DNA polymerase. *Biochem*. **30,** 7661-7666.

Noble, E.R. and **Noble, G.A**. (1982*). Parasitology: The biology of animal parasites Addison-Wesley Publishing. 5th Edition. pp.* 21-58.

North, M.J., Coombs, G.H. and **Barry, J.D**. (1983). A comparative study of the proteolytic enzymes of *Trypanosoma brucei, T. equiperdum, T. evansi, T. vivax, Leishmania tarentolae,* and *Crithidia fasciculata. Mol. Biochem. Parasitol.* **9**, 161-180.

Okhuysen, P.C., DuPont, H.L., Sterling, C.R. and Chappell, C.L. (1994). Arginine aminopeptidase, an integral membrane protein of the *Cryptosporidium parvum* sporozoite. *Infect. Immun*. **62,** 4667-4670.

Olivos-García, A., **Nequiz-Avendaño, M.**, **Tello, E.**, **Martínez, R.D.**, **González-Canto, A.**, **López-Vancell, R.**, **García de León, M.C.**, **Montfort, I. and Pérez-Tamayo, R**. (2004). Inflammation, complement, ischemia and amoebic survival in acute experimental amoebic liver abscesses in hamsters. *Exp. Mol. Pathol.* **77,** 66-71.

Omara-Opyene, A.L. and **Gedamu, L.** (1997). Molecular cloning, characterisation and overexpression of two distinct cysteine protease cDNA from *Leishmania donovani chagasi. Mol. Biochem. Parasitol.* **90,** 247-267.

Palomares, C.M., **Biteau, N.**, **Giroud, C.**, **Coustou, V.**, **Coetzer, T.**, **Authié, E.**, **Boulangé, A. and Baltz, T**. (2008). Molecular and biochemical characterization of a cathepsin B-like protease family unique to *Trypanosoma congolense. Eukaryot Cell.* **7**, 684-697.

Pamer, E.G., So, M. and Davis, C.E. (1989). Identification of a developmentally regulated cysteine protease of *Trypanosoma brucei. Mol. Biochem. Parasitol.* **33**, 27-32.

Pamer, E.G., Davis, C.E., Eakin, A. and So, M. (1990). Cloning and sequencing of the cysteine protease cDNA from *Trypanosoma brucei rhodesiense. Nucleic Acids Res*. **18,** 6141.

Pépin, J. and Méda, H.A. (2001). The epidemiology and control of human African trypanosomiasis. *Adv. Parasitol.* **49,** 71-132.

Pils, B. and Schultz, J. (2004). Inactive enzyme-homologues find new function in regulatory processes. *J. Mol. Biol.* **340,** 399-404.

Rawlings, N.D. and Barrett, A.J. (1994). Families of serine peptidases. *Methods Enzymol*. **244,** 19-61.

Rawlings, N.D. and Barrett, A.J. (1995a). Evolutionary families of metallopeptidases. *Methods Enzymol*. **248,** 183-228.

Rawlings, N.D. and Barrett, A.J. (1995b). Families of aspartic peptidases, and those of unknown catalytic mechanism. *Methods Enzymol*. **248,** 105-120.

Rawlings, N.D. and Barrett, A.J. (1999). MEROPS: the peptidase database. *Nucleic Acids Res*. **27,** 325-331.

Rawlings, N.D. and Barrett, A.J. (2000). MEROPS: the peptidase database. *Nucleic Acids Res*. **28,** 323-325.

Rebeski, D.E., Winger, E.M., Rogovic, B., Robinson, M.M., Crowther, J.R. and Dwinger, R.H. (1999). Improved methods for diagnosis of African trypanosomosis. *Memo. Inst. Oswaldo Cruz.* **94**, 249-253.

Richardson, C.C. (1965). Phosphorylation of nucleic acid by an enzyme from T4 bacteriophage-infected *Escherichia coli. Proc. Natl. Acad. Sci*. **54**, 158-165.

Rivett, A.J. (1993). Proteasomes: multicatalytic proteinase complexes. *J. Biochem*. **291,** 1-10.

Robertson, C.D. and Coombs, G.H. (1990). Characterisation of three groups of cysteine proteinases in the amastigotes of *Leishmania mexicana mexicana*. *Mol. Biochem. Parasitol.* **42**, 269-276.

Rosenthal, P.J. (1999). Proteases of protozoan parasites. *Adv. Parasitol.* **43**, 105-159.

Rosenthal, P.J. (2004). Cysteine proteases of malaria parasites. *Int. J. Parasitol*. **34,** 1489-1499.

Rosenthal, P.J. and Nelson, R.G. (1992). Isolation and characterization of a cysteine proteinase gene of *Plasmodium falciparum. Mol. Biochem. Parasitol.* **51,** 143-152.

Rosenthal, P.J., **Sijwali, P.S.**, **Singh, A. and Shenai, B.R**. (2002). Cysteine proteases of malaria parasites: targets for chemotherapy. *Curr. Pharm. Des.* **8,** 1659-1672.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. **and Arnheim, N.** (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anaemia. *Science*. **230**, 1350-1354.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with thermostable DNA polymerase. *Science*. **239**, 487-491.

Sajid, M. and McKerrow, J.H. (2002). Cysteine protease of parasitic organisms. *Mol. Biochem. Parasitol*. **120,** 1-21.

Sakanari, J.A., Nadler, S.A., Chan, V.J., Engel, J.C., Leptak, C. and **Bouvier, J.** (1997). *Leishmania major*: comparison of the cathepsin L- and B-like cysteine protease genes with those of other trypanosomatids. *Exp. Parasitol.* **85,** 63-76.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). Gel electrophoresis of DNA. Molecular cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY.

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

Schmidt, G.D. and Roberts, L.S. (1989). Foundations of parasitology: Times Mirror/Mosby College Publishing. *4th Edition pp* 55-80.

Schofield, C.J. and Maudlin, I. (2001). Trypanosomiasis control. *Int. J. Parasitol.* **31,** 614-619.

Seemüller, E., Lupas, A., Stock, D., Löwe, J., Huber, R. and Baumeister, W. (1995). Proteasome from *Thermoplasma acidophilum*: a threonine protease. *Science.* **268,** 579-582.

Serveau, C., Lalmanach, G., Hirata, I., Scharfstein, J., Juliano, M.A. and Gauthier, F. (1999). Discrimination of cruzipain, the major cysteine proteinase of *Trypanosoma cruzi*, and mammalian cathepsins B and L, by a pH-inducible fluorogenic substrate of trypanosomal cysteine proteinases. *Eur. J. Biochem*. **259,** 275-280.

Shenai, B.R., Sijwali, P.S., Singh, A. and Rosenthal, P.J. (2000). Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of *Plasmodium falciparum. J. Biol. Chem*. **275**, 29000-29010.

Sijwali, P.S., **Shenai, B.R.**, **Gut, J.**, **Singh, A**. **and Rosenthal, P.J.** (2001). Expression and characterization of the *Plasmodium falciparum* haemoglobinase falcipain-3. *Biochem. J.* **360,** 481-489.

Sijwali, P.S., **Kato, K.**, **Seydel, K.B.**, **Gut, J.**, **Lehman, J.**, **Klemba, M.**, **Goldberg, D.E.**, **Miller, L.H. and Rosenthal, P.J**. (2004). *Plasmodium falciparum* cysteine protease falcipain-1 is not essential in erythrocytic stage malaria parasites. *Proc. Natl. Acad. Sci.* **101,** 8721-8726.

Sijwali, P.S. and Rosenthal, P.J. (2004). Gene disruption confirms a critical role for the cysteine protease falcipain-2 in hemoglobin hydrolysis by *Plasmodium falciparum. Proc. Natl. Acad. Sci.* **101,** 4384-4389.

Sijwali, P.S., Koo, J., Singh, N. and Rosenthal, P.J. (2006). Gene disruptions demonstrate independent roles for the four falcipain cysteine proteases of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **150,** 96-106.

Simpson, A.G., Stevens, J.R. and Lukeš, J. (2006). The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol.* **22,** 168-174.

Singh, N., **Sijwali, P.S.**, **Pandey, K.C. and Rosenthal, P.J**. (2006). *Plasmodium falciparum*: biochemical characterization of the cysteine protease falcipain-2'. *Exp. Parasitol.* **112,** 187-192.

Souza, A.E., Waugh, S., Coombs, G.H. and Mottram, J.C. (1992). Characterization of a multi-copy gene for a major stage-specific cysteine proteinase of *Leishmania mexicana. FEBS Lett.* **311**, 124-127.

Souza, A.E., **Bates, P.A.**, **Coombs, G.H. and Mottram, J.C**. (1994). Null mutants for the *lmcpa* cysteine proteinase gene in *Leishmania mexicana. Mol. Biochem. Parasitol.* **63,** 213-220.

Stevens, J.R. and Gibson, W. (1999). The molecular evolution of trypanosomes. *Parasitol. Today*. **15,** 432-437.

Stevens, J.R., Noyes, H.A., Schofield, C.J. and Gibson, W. (2001). The molecular evolution of Trypanosomatidae. *Adv. Parasitol*. **48,** 1-56.

Steverding, D., Caffrey, C.R. and Sajid, M. (2006). Cysteine proteinase inhibitors as therapy for parasitic diseases: advances in inhibitor design. *Mini. Rev. Med. Chem*. **6,** 1025-1032.

Stoka, V., McKerrow, J.H., Cazzulo, J.J. and Turk, V (1998). Substrate inhibition of cruzipain is not affected by the C-terminal domain. *FEBS Lett.* **429**, 129-133.

Stoka, V., Turk, B., McKerrow, J.H., Björk, I., Cazzulo, J.J. and Turk, V. (2000). The high stability of cruzipain against pH-induced inactivation is not dependent on its C-terminal domain. *FEBS Lett.* **469**, 29-32.

Storer, A.C. and Ménard, R. (1994). Catalytic mechanism in papain family of cysteine peptidases. *Methods Enzymol*. **244,** 486-500.

Stuart, K.., Allen, T.E., Heidmann, S. and Seiwert, S.D. (1997). RNA editing in kinetoplastid protozoa. *Microbiol. Mol. Biol. Rev*. **6,** 105-120.

Thompson J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22, 4673-4680.

Tomás, A.M. and Kelly, J.M. (1996). Stage-regulated expression of cruzipain, the major cysteine protease of *Trypanosoma cruzi* is independent of the level of RNAI. *Mol. Biochem. Parasitol.* **76**, 91-103.

Traub-Cseko, Y.M., Duboise, M., Boukai, L.K. and McMahon-Pratt, D. (1993). Identification of two distinct cysteine proteinase genes of *Leishmania pifanoi* axenic amastigotes using the polymerase chain reaction. *Mol. Biochem. Parasitol.* **57,** 101-115.

Travis, J. and Potempa, J. (2000). Bacterial proteinases as targets for the development of second-generation antibiotics. *Biochim. Biophys. Acta*. **1477,** 35-50.

Troeberg, L., Pike, R.N., Morty, R.E., Berry, R.K., Coetzer, T.H. and Lonsdale-Eccles, J. D. (1996). Proteases from *Trypanosoma brucei brucei*. Purification, characterisation and interactions with host regulatory molecules. *Eur. J. Biochem.* **238**, 728-736.

Troeberg, L., Morty, R.E., Pike, R.N., Lonsdale-Eccles, J.D., Palmer, J.T., McKerrow. JH. and Coetzer T.H. (1999). Cysteine proteinase inhibitors kill cultured bloodstream forms of *Trypanosoma brucei brucei*. *Exp. Parasitol*. **91,** 349-355.

Turner, C.M., Melville, S.E. and Tait, E. (1997). A proposal for karyotype nomenclature in *Trypanosoma brucei*. *Parasitol. Today*. **13,** 5-6.

Van den Bossche, P., Chigoma, D. and Shumbwa, W. (2000). The decline of anti- trypanosomal antibody levels in cattle after treatment with trypanocidal drugs and in the absence of the tsetse challenge. *Acta trop.* **77**, 263-270.

Vickerman, K. (1969). The fine structure of *Trypanosoma congolense* in its bloodstream phase. *J. Protozool*. **16,** 54-69.

Wang, C., Barry, J.K., Min, Z., Tordsen, A., Rao, A.G. and Olsen, O.A. (2003). The calpain domain of the maize DEK1 protein contains the conserved catalytic triad and functions as a cysteine proteinase. *J. Biol. Chem*. **278**, 34467-34474.

Weiden, M., Osheim, Y.N., Beyer, A.L. and Van der Ploeg, L.H. (1991). Chromosome structure: DNA nucleotide sequence elements of a subset of the mini-chromosomes of the protozoan *Trypanosoma brucei*. *Mol. Cell. Biol.* **11**, 3823-3834.

Weiss, B., Jacquemin-Sablon, A., Live, T.R., Fareed, G.C. and Richardson, C.C. (1968). Enzymatic breakage and joining of deoxyribonucleic acids. VI. Further purification and properties of polynucleotide ligase from *Escherichia coli* infected with bacteriophage T4. *J. Biol. Chem*. **243**, 4543-4555.

Welburn, S.C., Picozzi, K., Fèvre, E.M., Coleman, P.G., Odiit, M., Carrington, M. and Maudlin, I. (2001). Identification of human-infective trypanosomes in animal reservoir of sleeping sickness in Uganda by means of serum-resistance-associated (SRA) gene. *Lancet*. **358**, 2017-2019.

Williams, D.J., Taylor, K., Newson, J., Gichuki, B. and Naessens, J. (1996). The role of anti-variable surface glycoprotein antibody responses in bovine trypanotolerance. *Parasite Immunol*. **18,** 209-218.

Williams, J.F. (1989). Optimization strategies for the polymerase chain reaction. *Biotechniques*. **7**, 762-769.

Williamson, C.C., Stoltsz, W.H, Mattheus, A. and Schiele, G.J. (1988). An investigation into alternative methods for the serodiagnosis of dourine. Onderstepoort. *J. Vet. Res*. **55,** 117-119.

Yabu, Y., Suzuki, T., Nihei, C., Minagawa, N., Hosokawa, T., Nagai, K., Kita, K. and Ohta, N. (2006). Chemotherapeutic efficacy of ascofuranone in *Trypanosoma vivax*-infected mice without glycerol. *Parasitol. Int*. **55,** 39-43.

Zadeh-Vakili, A., Taheri, T., Taslimi, Y., Doustdari, F., Salmanian, A.H. and Rafati, S. (2004). Immunization with the hybrid protein vaccine, consisting of *Leishmania major* cysteine proteinases Type I (CPB) and Type II (CPA), partially protects against leishmaniasis. *Vaccine.* **22,** 1930-1940.

Zhang, M., Wei, Z., Chang, S., Teng, M. and **Gong, W.** (2006). Crystal structure of a papain- fold protein without the catalytic residue: a novel member in the cysteine proteinase family. *J. Mol. Biol*. **358**, 97-105.

Zhang, Z., **Wang, L.**, **Seydel, K.B.**, **Li, E.**, **Ankri, S.**, **Mirelman, D.**, **Stanley, S.L. Jr**. (2000). *Entamoeba histolytica* cysteine proteinases with interleukin-1 beta converting enzyme (ICE) activity cause intestinal inflammation and tissue damage in amoebiasis. *Mol. Microbiol.* **37,** 542-548.

APPENDIX A: The restriction map of the congopain gene
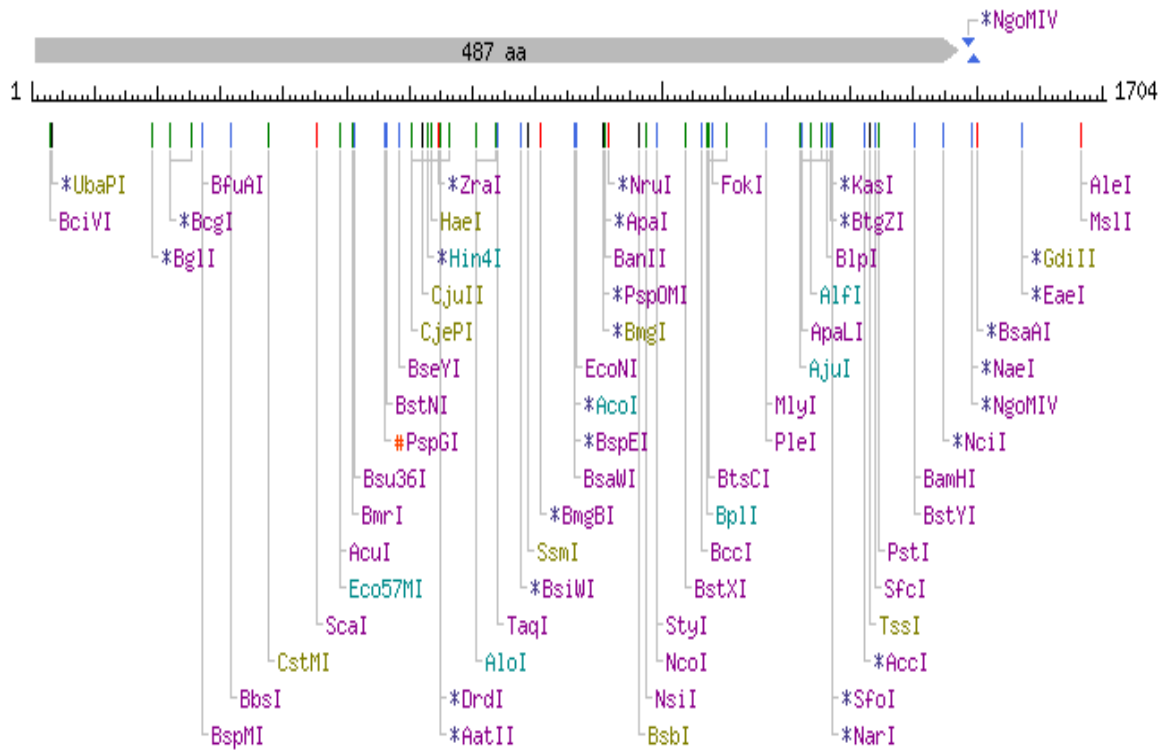
APPENDIX B: Design of degenerate primers

The primer CP1 targeted the region coding for FKQKYS in the propeptide, while CP2 targeted (AEEEAFRFR) a conserved region in the pro-region of cysteine proteases in trypanosomatids. CP3 targeted (SEQMLV) a conserved region in catalytic domain of cysteine proteases among trypanosomes species. The region targeted by CP4 primer (WGEEGYIRI) occurs between the asparagine residue of the catalytic triad and the polyproline hinge. The region targeted by CP6 primer codes for SSSCSG in the C-terminal extension.

```
                         (CP1)                                          (CP2)
T.congoCP1    TTCGCCGCA TTCAAGCAAAAGTACAGCA GGTCGTACAAGGAC GCCACGGAGGAGGCATTC 180
T.congoCP2    TTCGCCGCA TTCAAGCAAAAGTACAGCA GGTCGTACAAGGAC GCCACGGAGGAGGCATTC 180
T.bruceiCP    TTTGCTGCG TTCAAGAAGAAGTACGGCA AGGTGTACAAGGAT GCTAAGGAGGAAGCATTC 180
T.cruziCP1    TTCGCAGAA TTCAAGCAGAAGCATGGCA GGGTGTACAGAGC GCCGCGGAGGAGGCGTTC 171
T.cruziCP2    TTCGCAGAA TTCAAGCAGAAGCATGGCA GGGTGTACGGGAGC GCCGCGGAGGAGGCGTTC 171
              ** ** *    ****** * *** * ***   *  ****  *    **   ****** ** ***


T.congoCP1    CGTTTCCG CGTCTTCAAGCAGAACATGGAGCGTGCAAAGGAGGAGGCCGCTGCGAACCCC 240
T.congoCP2    CGTTTCCG TGTCTTCAAGCAGAACATGGAGCGTGCAAAGGAGGAGGCCGCTGCGAACCCC 240
T.bruceiCP    CGCTTCCG TGCCTTTGAGGAAAATATGGAGCAGGCGAAGATTCAAGCTGCGGCGAACCCA 240
T.cruziCP1    CGCCTGAG CGTGTTCAGGGAGAACCTGTTTCTTGCGAGGCTGCACGCCGCGGCAAACCCA 231
T.cruziCP2    CGCCTGAG CGTGTTCCGGGCGAACCTGTTTCTTGCGAGGCTGCACGCCGCGGCAAACCCA 231
              **   *   *   *  **    *    ** **    *  ** * *    *  ** ** ** *****


(CP3)                                              (CP3)
T.congoCP1    GGCCAGTGGAAGGTTGCGGGCCATGAGCTGACGTCTCTG TCGGAGCAAATGCTCGTGTC A 540
T.congoCP2    GGCCAGTGGAAGGTTGCAGGCCATGAGCTGACGTCTTTG TCGGAGCAGATGCTCGTGTC A 540
T.bruceiCP    GGGCAGTGGCAGGTGGCAGGAAATCCTCTCGTATCCCTC TCGGAGCAGATGCTAGTGTC A 540
T.cruziCP1    TGCCAGTGGTTTCTTGCCGGCCACCCGCTGACGAACCTG TCGGAGCAGATGCTCGTGTC G 531
T.cruziCP2    TGCCAGTGGTTTCTCGCCGGCCACCCGCTGACGAACCTG TCGGAGCAGATGCTCGTGTC G 531
               * ******    * ** **  *     **      *   ******** ***** *****
                                                   S   E   Q   M   L   V   S

                                          (CP4)
T.congoCP1    CCATATTGGATTATCAAGAACTCATGGAGCAAGGGA TGGGGCGAGGAGGGTTACATCCGC 960
T.congoCP2    CCATACTGGATTATCAAGAACTCATGGAGCAAGGGA TGGGGCGAGGAGGGTTAC-TCCGC 959
T.bruceiCP    CCCTACTGGATCATCAAAAACTCGTGGAGCAACATG TGGGGCGAGGACGGCTACATCCGC 960
T.cruziCP1    CCGTACTGGATCATCAAGAACTCGTGGACCACGCAG TGGGGCGAGGAAGGCTACATCCGC 951
T.cruziCP2    CCGTACTGGGTCATCAAGAACTCGTGGACCACGCAG TGGGGTGAGGATGGCTACATCCGC 951
              ** ** *** * ***** ***** **** **     ***** ***** ** *** *****


T.congoCP1    AT TGAGAAGGGCACAAACCAATGTCTCATGAAAAATTACGCGAGGTCAGCCGTTGTCAGC 1020
T.congoCP2    AT TGAGAAGG-CACAA-CCAATGTCTTATGAAAATCTCCCAAGCTCCGCCGTTGTCAGC 1017
T.bruceiCP    AT CGAGAAGGGCACAAACCAATGTCTCATGAATCAGGCCGTATCCTCCGCAGTTGTTGGA 1020
T.cruziCP1    AT TGCAAAGGGCTCGAACCAGTGCCTTGTCAAGGAGGAGGCGAGCTCCGCGGTGGTCGGT 1011
T.cruziCP2    AT TGCAAAGGGCTCGAACCAGTGCCTTGTGAAGGAGGAGGCGAGCTCCGCTGCCAAGCAT 1011
              **   *  **** * * * *** ** ** * **  *       ** ** *
              I

                                          (CP6)
T.congoCP1    TGCGGCTCCAATAACCTCACACAGATCGTCTACCCGTTG AGCAGCTCCTGCAGCGG CTTC 1227
T.congoCP2    TGCGGCTCCAATAACCTCACGCAGATCGTCTACCCGTTG AGCAGCTCCTGCAGCGG CTTC 1218
T.bruceiCP    TGTGGCGCAAGCAACCTTACACAAATAATCTACCCACTA AGCAGGAGCTGCAGCGG TCCC 1236
T.cruziCP1    TGCGGTGCTGAGACTCTCACAGAAGAAGTCTTCCTTACG AGTACGCACTGCAGCGG CCCA 1251
T.cruziCP2    --------------------------------------- ----------------- ----
```

APPENDIX C: Design of *T. congolense* specific primer Tc/Tv CP2/CP3

The primer Tc/Tv CP2 targeted MEGAKEE a conserved region in the propeptide, Tc/Tv CP3 targeted
SEQMLVS a conserved region at the catalytic domain

```
CLUSTAL W (1.83) multiple sequence alignment

                                                          ───────────────▶ (Tc/Tv CP2)
Tcs62    GGAGGAGGCGTTCCGTTTCCGCGTCTTCAAGCAGAA CATGGAGCGTGCAAAGGAGGAGG C 60
Tcs66    GGAGGAGGCGTTCCGTTTCCGTGTCTTCAAGCAGAA CATGGAGCGTGCAAAGGAGGAGG C 60
Tcs58    GGAGGAGGCGTTCCGTTTCCGCGTCTTCAAGCAGAA CATGGAGCGTGCAAAGGAGGAGG C 60
Tcs41    GGAGGAGGCGTTCCGTTTCCGCTTCTTCAAGCAGAA CATGGAGCGTGCAAAGGAGGAGG C 60
Tcs59    GGAGGAGGCGTTCCGTTTCCGCGTCTTCAAGCAGAA CATGGAGCGTGCAAAGGAGGAGG C 60
Tcs65    GGAGGAGGCGTTCCGTTTCCGCGTCTTCAAGCAGAA CATGGAGCGTGCAAAGGAGGAGG C 60
         *********************  ************ ********************* *


Tcs62    CGCTGCGAACCCCTATGCGACGTTTGGTGTGACGCGGTTCTCCGATATGCCACCCGAGGA 120
Tcs66    CGCTGCGAACCCCTATGCGACGTTTGGTGTGACGCGGTTCTCCGATATGTCACCCGAGGA 120
Tcs58    CGCTGCGAACCCCTATGCGACGTTTGGTGTGACGCGGTTCTCCGATATGTCACCCGAGGA 120
Tcs41    CGCTGCGAACCCCTATGCGACGTTTGGTGTGACGCGGTTCTCCGATATGTCACCCGAGGA 120
Tcs59    CGCTGCGAACCCCTATGCGACGTTTGGTGTGACGCGGTTCTCCGATGTGTCACCCGAGGA 120
Tcs65    CGCTGCGAACCCCTATGCGACGTTTGGTGTGACGCGGTTCTCCGATATGTCACCCGAGGA 120
         ***********************************************  ** *********


Tcs62    GTTCAGGGCGACCTACCACAACGGGGCGGAGTACTACGCTGCGGCGCTGAAGCGACCACG 180
Tcs66    GTTCAGGGCGACCTACCACAACGGGGCGGAGTACTACGCTGCGGCGCTGAAGCGACCACG 180
Tcs58    GTTCAGGGCGACCTACCACAACGGGGCGGAGTACTACGCTGCGGCGCTGAAGCGACCACG 180
Tcs41    GTTCAGGGCGACCTACCACAACGGGGCGGAGTACTACGCTGCGGCGCTGAAGCGACCACG 180
Tcs59    GTTCAGGGCGACCTACCACAACGGGGCGGAGTACTACGCTGCGGCGCTGAAGCGACCACG 180
Tcs65    GTTCAGGGCGACCTACCACAACGGGGCGGAGTACTACGCTGCGGCGCTGAAGCGACCACG 180
         ************************************************************


Tcs62    CAAGGTGGTGAATGTGTCCACTGGGAAGGCACCTGAGGCAGTTGACTGGCGCAAGAAGGG 240
Tcs66    CAAGGTGGTGAATGTGTCCACTGGGAAGGCACCTGAAGCAGTTGACTGGCGCAAGAAGGG 240
Tcs58    CAAGGTGGTGAATGTGTCCACTGGGAAGGCACCTGAGGCAGTTGACTGGCGCAAGAAGGG 240
Tcs41    CAAGGTGGTGAATGTGTCCACTGGGAAGGCACCTGAGGCAGTTGACTGGCGCAAGAAGGG 240
Tcs59    CAAGGTGGTGAATGTGTCCACTGGGAAGGCACCGCCGGCAATTGACTGGCGCAAGAAGGG 240
Tcs65    CAAGGTGGTGAATGTGTCCACTGGGGAGGCACCGCCGGCAATTGACTGGCGCAAGAAGGG 240
         ************************* *******    *** ******************


Tcs62    CGCGGTGACACCCGTGAAGGACCAGGGG-CGATGCGGCTCTTGCTGGGCATTCTCTGCCA 299
Tcs66    CGCGGTGACACCCGTGAAGGACCAGGGG-CGATGCGGCTCTTGCTGGGCATTCTCTGCCA 299
Tcs58    CGCGGTGACACCCGTGAAGGACCAGGGA-CAATGCGGTTCTTGCTGGGCATTCTCTGCCA 299
Tcs41    CGCGGTGACACCCGTAAAGGACCAGGGG-AAATGCGACTCTTCGTGGGCATTCACTGTCA 299
Tcs59    CGCGGTGACACCCGTGAAGGACCAGGGG-CAATGCGGCTCTTGCTGGGCATTCTCTGCCA 299
Tcs65    CGCGGTGACACCCGTGAAGGACGAGGGGGAAATGCGACTCTTTTTGGGCATTCTCTGCC- 299
         *************** ****** ****     *****  ****  ********* *** *


Tcs62    TAGGGAACATAGA-GGGCCAGTGGAAGATTGCGGGCCATGAGCTGACGTCTCTG TCGGAG 358
Tcs66    CAGGGAACATAGA-GGGCCAGTGGAAGATTGCGGGCCATGAGCTGACGTCTCTG TCGGAG 358
Tcs58    TAGGGAACATAGA-GGGCCAGTGGAAGATTGCGGGCTATGAGCTGACGTCTCTG TCGGAG 358
Tcs41    TAGGGAACATAGA-GGGTCAGTGGAAGATTGCGGGCCATGAGCTGACGTCTCTG TCGGAG 358
Tcs59    TAGGGAACATAGA-GGGCCAGTGGAAGATTGCGGGCCATGAGCCGACGTCTTTG TCGGAG 358
Tcs65    TAGGGAACATAGAAGGGTCAGTGGAAGATTGCGGGCCATGAGCTGACGTCTCTG TCGGAG 359
         ************ *** ***************** ****** ******* ** ******


                   ◀───────────── (Tc/Tv CP3)
Tcs62    CAAATGCTAGTGTC AAT 375
Tcs66    CAAATGCTAGTGTC AAT 375
Tcs58    CAGATGCTAGTGTC AAT 375
Tcs41    CAAATGCTAGTGTC AAT 375
Tcs59    CAAATGCTAGTGTC AAT 375
Tcs65    CAGATGCTAGTGTC AAT 376
         ** ********** ***
```

APPENDIX D: Design of *T. congolense* specific primers Tc CP1/CP4

The primer Tc CP1 targeted VIASCGS in the C-terminal extension; while Tc CP4 targeted
CAAAAATAAAAAGACAAGGCCAACGGG of intergenic region.

```
CLUSTAL W (1.83) multiple sequence alignment


Tcs57     AGTAAGTGCACCAAAGCCACATTCCCAACGGGCAAGTGCGTGCAGCTCAGCGGCGCCGGT 60
Tcs58     AGTAAGTGCACCAAAGCCACATTCCCAACGGGCAAGTGCGTGCAGCTCAGCGGCGCCGGT 60
Tcs66     AGTAAGTGCACCAAAGCCACATTCCCAACGGGCAAGTGCGTGCAGCTCAGCGGCGCCGGT 60
Tcs49     AGTAAGTGCACCAAAGCCACATTCCCAACGGGCAAGTGCGTGCAGCTCAGCGACACCGGC 60
          *********************************************** * ****
                         ───────────────▶ (Tc CP1)
Tcs57     TCT GTCATCGCCTCGTGCGGCTCC AATAACCTCACACAGATCGTCTACCCGTTGAGCAGC 120
Tcs58     TCT GTCATCGCCTCGTGCGGCTCC AATAACCTCACACAGATCGTCTACCCGTTGAGCAGC 120
Tcs66     TCT GTCATCGCCTCGTGCGGCTCC AATAACCTCACGCAGATCGTCTACCCGTTGAGCAGC 120
Tcs49     TCC GCCATCGCCTCGTGCGGCTCC AATAACCTCACACAGATCGTCTACCCGTTGAGCAGC 120
          **   * ******************* *********** ***********************


Tcs57     TCCTGCAGCGGCTTCTCCGTTCCGTTGACTGTGCCACTGGACAAGTGCCTGCCCATCGTG 180
Tcs58     TCCTGCAGCGGCTTCTCCGTTCCGTTGACTGTGCCACTGGACAAGTGCCTGCCCATCGTG 180
Tcs66     TCCTGCAGCGGCTTCTCCGTTCCGTTGACTGTGCCACTGGACAAGTGCCTGCCCATCGTG 180
Tcs49     TCCTGCAGCGGCTTCTCCGTTCCGTTGACTGTGCCACTGGACAAGTGCCTGCCCATCGTG 180
          ************************************************************


Tcs57     ATTGGATCCGTGATGTATGAGTGCTCTGACAGGGTTCCTACGGAATCCGCCCGGCTCGTG 240
Tcs58     ATTGGATCCGTGATGTATGAGTGCTCTGACAAGGCTCCTACGGAATCCGCCCGGCTCGTG 240
Tcs66     ATTGGATCCATGATGTACAAGTGCTCTGACAAGGCTCCTACGGAATCCGCCCGGCTCGTG 240
Tcs49     ATTGGATCCATGATGTATGAGTGCTCTGACAAGGCTCCTACGGAATCCGCCCGGCTCGTG 240
          ********* *******  *********** ** ***********************

CLUSTAL W (1.83) multiple sequence alignment

                               (Tc CP4)◀───────────────
Tcs49     GTATTGGGAGCAAGTGTGGGTG CAAAAATAAAAAGACAAGGCCAACGGG GTGCGACGCTC 60
Tcs57     GTATTGGGAGCAAGTGTGGGTG CAAAAATAAAAAGACAAGGCCAACGGG GTGCGACGCTC 60
Tcs66     GTATTGGGAGCAAGTGTGGGTG CAAAAATAAAAAGACAAGGCCAACGGG GTGCGACACTC 60
Tcs58     GTATTGGGAGCAAGTGTGGGTG CAAAAATAAAAAGACAAGGCCAACGGG GTGCGACGCTC 60
          ********************* *************************** ******* ***


Tcs49     CTCGCGGGTCAGTGGCAGCAGCGACTGTTGCCGTGGCC 98
Tcs57     CTCGCGGGTCAGTGGCAGCAGCGACTGTTGCCGTGGCC 98
Tcs66     CTCGCGGGTCAGTGGCAGCAGCGACTGTTGCCGTGGCC 98
Tcs58     CTCGCGGGTCAGTGGTAGCAGCGACTGTTGCCGTGGCC 98
          *************** *********************
```