Specific *cpb* copies within the *Leishmania donovani* complex: evolutionary interpretations and potential clinical implications in humans

M. HIDE[1]*, R. BRAS-GONÇALVES[2] *and* A. L. BAÑULS[1]

[1] *Génétique et Evolution des Maladies Infectieuses, IRD/CNRS (UMR 2724), F-34394, France*
[2] *IRD de Montpellier, Pathogénie des trypanosomatidés, UR08, 911, avenue Agropolis BP 64501, 34394 Montpellier Cedex 5, France*

(*Received 12 April 2006; revised 17 June 2006; accepted 22 August 2006; first published online 28 November 2006*)

SUMMARY

Leishmania infantum and *Leishmania donovani* both pertain to the *L.* (*L.*) *donovani* complex and are responsible for visceral leishmaniasis. To explore the *L. donovani* complex, we focused our study on *cysteine protease B* (*cpb*) and especially on 2 *cpb* copies: *cpbE* and *cpbF*. We selected *cpb* genes because of their phylogenetic interest and host–parasite interaction involvement. Sequencing these 2 copies revealed (i) that *cpbE* is specific to *L. infantum* and *cpbF* is specific to *L. donovani* and (ii) that these 2 copies are different in length and sequence.† Phylogenetic analysis and protein predictions were carried out in order to compare these copies (i) with other trypanosomatid *cpb*, especially *L. mexicana*, and (ii) within the *L. donovani* complex. Our results revealed patterns specific to the *L. donovani* complex such as the COOH-terminal extension, potential epitopes and N-glycosylation sites. Moreover, phylogenetic analysis revealed different levels of polymorphism between *L. infantum* and *L. donovani* and confirmed the ancestral status of the latter. *L. infantum* has a shorter sequence and a deleted sequence responsible for modifications in protein conformation and catalytic triad. Considering the clinical aspect, *L. infantum* dermotropic strains appeared more polymorphic than *L. infantum* viscerotropic strains.

Key words: cysteine protease B, *Leishmania donovani* complex, phylogeny, protein isoforms, COOH-terminal extension.

INTRODUCTION

Leishmaniasis, caused by the *Leishmania* parasite, is a worldwide endemic disease, with an estimated disease burden of 2 357 000 disability-adjusted life years and 59 000 deaths per year (WHO, 2002). This disease ranges in severity from a healing skin ulcer to an overwhelming visceral form. The visceral form, the most severe form, is caused by the species of the *Leishmania donovani* complex, and principally by *Leishmania infantum* and *L. donovani*. These species are associated with different epidemiology, ecology and pathology: *L. infantum* is anthropozoonotic with a dog reservoir and can produce visceral and cutaneous forms in humans, whereas *L. donovani* is largely anthroponotic and produces mainly the visceral form.

* Corresponding author: Laboratory GEMI, UMR CNRS/IRD 2724, 911, avenue Agropolis BP 64501, 34394 Montpellier Cedex 5, France. Tel: +33 4 67 41 62 72. Fax: +33 4 67 41 62 99. E-mail: hide@mpl.ird.fr

† Nucleotide sequence data reported in this paper are available in the GenBank database under Accession numbers AY896776 AY896777, AY896778, AY896779, AY896780, AY896781, AY896782, AY896783, AY896784, AY896785, AY896786, AY896787, AY896788, AY896789, AY896790, AY896791.

Several pathogenic protozoan parasites, like *Leishmania*, express multiple cysteine protease (CP) enzymes. These enzymes pertain to the papain superfamily and within the *Leishmania* genus, 3 types of CPs have been revealed: CPA and CPB, which have homology to mammalian cathepsin L, and CPC, which has homology to mammalian cathepsin B (Mottram *et al*. 1997, 1998; Sajid and McKerrow, 2002). Cysteine proteases are important for *Leishmania* survival, host cell infection, and evasion of the host immune response, and they have attracted considerable interest as targets for the design of new chemotherapy (Alexander *et al*. 1998; Beyrodt *et al*. 1997; Coombs and Mottram, 1997; Frame *et al*. 2000; McGrath *et al*. 1995; Mottram *et al*. 1996; Sajid and McKerrow, 2002; Souza *et al*. 1992). Furthermore, cysteine proteases are putative virulence factors of *Leishmania* parasites. *Cpa* and *cpc* are single-copy genes, whereas *cpb* is a multicopy gene. Differences in copy number and nucleotide sequences exist among the different *Leishmania* species. For example, *Leishmania mexicana cpb*, the most studied, are located in a single locus of 19 copies arranged in a tandem repeat (Mottram *et al*. 1996, 1997), whereas the *L. donovani* complex seems to be composed of 5 copies. Indeed, Mundodi *et al*. (2002) compared 1 *L. donovani* strain and 1 *L. chagasi*

doi:10.1017/S0031182006001600 Printed in the United Kingdom

Table 1. *Leishmania* strains used in this study

	Species	WHO code	Zymodeme[a]	Clinical data[b]	GenBankAC no.[c]
L. donovani complex	*L. donovani*	MHOM/ET/67/HU3	MON-18	VL	*cpbF* AY896783
	L. donovani	MHOM/ET/00/HUSSEN	LON-42	VL	*cpbF* AY896785
	L. donovani	MHOM/KE/67/MRC(L)3	LON-44 or 46	NC	*cpbF* AY896786
	L. donovani	IMAR/KE/62/LRC-L57	MON-37	—	*cpbF* AY896788
	L. donovani	MHOM/SD/82/GILANI	MON-30	PKDL	*cpbF* AY896784
	L. donovani	MHOM/CN/00/WangJie1	MON-35	VL	*cpbF* AY896787
	L. infantum	MHOM/FR/78/LEM75	MON-1	VL	*cpbE* AY896781
	L. infantum	MHOM/FR/82/LEM356	MON-33	CL	*cpbE* AY896779
	L. infantum	MHOM/FR/87/LEM1098	MON-1	CL	*cpbE* AY896776
	L. infantum	MHOM/FR/85/LEM716	MON-1	VL	*cpbE* AY896777
	L. infantum	MHOM/FR/85/LEM663	MON-1	VL	*cpbE* AY896780
	L. infantum	MHOM/ES/81/BCN1	MON-29	CL	*cpbE* AY896778
	L. infantum	MHOM/MA/67/ITMAP263	MON-1	VL	*cpbE* AY896782
	L. infantum	MHOM/DZ/82/LIPA59	MON-24	CL	*cpbE* AY896790 *cpbF* AY896789
	L. chagasi	MHOM/PA/78/WR285	MON-1	CL	*cpbE* AY896791

[a] Zymodemes were determined with the MLEE method performed by the WHO reference centres of Montpellier (MON) and London (LON).
[b] Clinical forms are noted: visceral leishmaniasis (VL), cutaneous leishmaniasis (CL), Post Kala-Azar dermal leishmaniasis (PKDL) and unknown (NC).
[c] The *cpbE* and *cpbF* sequences submitted to GenBank.

(syn *L. infantum*) strain and revealed at least 5 tandemly arranged genes. They observed that the *cpb* cluster of the *L. donovani* strain contains a *cpbF* copy which is absent from the cluster of the *L. infantum* strain. Furthermore, another copy called *cpbE* is distant from the cluster for the *L. donovani* strain, whereas it belongs to the cluster of the *L. infantum* strain (Mundodi *et al*. 2002). To explore the *L. donovani* complex, we focused our study on *cpbE* and *cpbF*. We sequenced them for a sample representative of the clinical (strains isolated from cutaneous leishmaniasis and visceral leishmaniasis) and genetic diversity of *L. donovani*/*L. infantum*. Phylogenetic analysis and protein predictions were conducted in order to compare these copies with other trypanosomatid *cpb* and those within the *L. donovani* complex. Evolutionary interpretations and potential clinical implications are discussed.

MATERIALS AND METHODS

Parasites

A sample of 15 strains representative of geographical and genetic diversity within the *L. donovani* complex was studied (Table 1). The 15 strains were isolated either from visceral or cutaneous human leishmaniasis or from phlebotomine sand fly and pertain to either the *L. donovani* species (6 strains) or to the *L. infantum* species (9 strains). All the *Leishmania* strains were typed by MLEE by the WHO reference centres of Montpellier, France (MON) and London (LON).

Cell culture

Promastigote cultures were maintained at 26 °C by weekly subpassages in RPMI 1640 medium, buffered with 25 mM HEPES, 2 mM NaHCO3 and supplemented with 10% heat-inactivated fetal calf serum, 2 mM glutamine, 100 U/ml penicillin and 100 μg/ml streptomycin. Cultures were harvested by centrifugation and stored at −80 °C until DNA extraction.

Specific PCR of cpbEF *copies*

Genomic DNA was extracted from parasite pellets by phenol/chloroform extraction. The optimal conditions for *cpbEF* amplification in 30 μl were: 6 pmol of each primer (forward: 5′-CGTGACGCCGGTGAAGAAT-3′; reverse: 5′-CGTGCACTCGGCCGTCTT-3′), 4·5 nmol dNTPs, 1 U Taq polymerase (Roche Diagnostics), 3 μl of Buffer 10X and 10 ng of genomic DNA. Thirty cycles were necessary for amplification (denaturation 30 s at 94 °C, annealing 1 min at 62 °C and elongation 1 min at 72 °C) followed by 10 min at 72 °C. The amplification reactions were analysed by agarose gel electrophoresis, followed by ethidium bromide staining and visualization under UV light. Two lengths of amplification products were generated: 702 bp for *cpbE* and 741 bp for *cpbF*.

Cloning and sequencing

For the 15 strains of the *L. donovani* complex, the PCR products were cloned into pGEM-T vector

(pGEM-T Easy Vector System I, Promega) and transformed into competent *E. coli* (JM109). Nucleotide sequences were obtained by automated sequencing (ABI PRISM™ 310 Genetic Analyzer, Applied Biosystems) and chromatograms were analysed with Chromas 2.23 (Technelysium Pty Ltd, 1998–2002).

Sequence analysis

For phylogenetic analysis, our *cpbE/cpbF* sequences were compared with 4 available sequences of *L. donovani* (GenBank Accession no. AF309627), *L. infantum* (AF217087 and AJ628943) and *L. major cpb* (U43706). Nucleotide and protein sequences were aligned using the Multiple Sequence Alignment Program, ClustalX (version 1.81, June 2000) (Thompson *et al*. 1997). Phenetic and phylogenetic analyses were performed with the PHYLIP package: we used distance methods with DNADIST, PROTDIST, NEIGHBOR programs and parsimony methods with SEQBOOT, DNAPARS, PROTPARS and CONSENSE programs. Concerning the phenetic analysis, various distance types were used to build the trees: Kimura two-parameter (Kimura, 1980), Jukes-Cantor (Jukes and Cantor, 1969) and Maximum Likelihood (Felsenstein, 1981). For parsimony analysis, the bootstrap analyses were performed for 1000 replications to estimate the robustness of the nodes. All the trees were constructed using TreeDyn software ((Chevenet *et al*. 2006), http://www.treedyn.org). The genotypic diversity rate (for either *cpbE* or *cpbF* sequences) was obtained by dividing the number of divergent sequences with the total number of sequences analysed.

The DNA sequences were also analysed considering the mutation sites and several available sequences were used to compare CPBEF sequences with other CPBs of *Leishmania* such as *L. chagasi ldccys1* (GenBank Accession no. AF004592), *L. mexicana lmcpbr* (Z14061), *L. mexicana cpb18* (Y09958), *L. mexicana lmcpb2.8* (Z49962), *L. mexicana cpb1* (Z49963), *L. mexicana cpb2* (AJ319727) and *L. mexicana cpb19* (Z49965).

Predicted protein structure

The impact of amino acid mutations on the 3D protein structure was studied with Swiss-Model (swissmodel.expasy.org/spdbv). The crystal structure of cruzain (McGrath *et al*. 1995) was used as the reference structure. Identity with N-glycosylation sites was checked using NetGlyc1.0 (www.cbs.dtu.dk/services/NetNGlyc). Predicted MHC Class-I and II binding regions (T-cell epitopes) were analysed using web servers: MAPPP (MHC-1 Antigenic Peptide Processing Prediction – www.mpiib-berlin.mpg.de/MAPPP), SYFPEITHI (www.syfpeithi.de), RANKPEP (mif.dfci.harvard.edu/Tools/rankpep.html), and ProPred (www.imtech.res.in/raghava/propred/). Concerning B-cell epitopes, prediction of antigenic peptides was based on amino acid residues in experimentally known segmental epitopes (Prediction Antigenic Peptides – www.mifoundation.org/Tools/, ABCpred – www.imtech.res.in/raghava/abcpred/index.html) and on antigenic peptides, which should be located in solvent-accessible regions identified from 3D structures.

RESULTS

PCR products were sequenced for the entire sample (15 strains). All sequences were submitted to GenBank™ (see GenBank Accession numbers in Table 1). For the 6 *L. donovani* strains, only the *cpbF* copy was obtained, whereas the 9 *L. infantum* strains contained only *cpbE*, except 1 strain. Indeed, this strain, LIPA59 from Algeria and isolated from the cutaneous form, had a mixed pattern comprising the 2 *cpb* copies *E* and *F*. This pattern (*E* and *F*) has been obtained from different cultures of a given LIPA59 clone (obtained by micromanipulation) and also from different clones. Thus, this pattern could not be due to a mixture of 2 strains or just to a contamination. Consequently, 16 sequences were obtained in this study. Other DNA samples from the *L. donovani* complex (*L. archibaldi*, *L. donovani* from India and *L. infantum* from Greece and Tunisia) have been amplified (data not shown) and gave similar results: *L. infantum* only contained the *cpbE* copy whereas *L. donovani* and *L. archibaldi* contained the *cpbF* copy. Their amplification products have not been sequenced.

Phylogenetic analysis

To obtain a comprehensive view of *cpbEF* polymorphism within the *L. donovani* complex, we present here the dendrogram constructed using the parsimony method on DNA sequences (Fig. 1). See URL (http://www.treedyn.org/hide/hide2005b.html) containing the coloured dendrogram and a direct link to GenBank. The 2 sequences of *L. infantum* LIPA59 were excluded because their lengths were too short, and the gaps in 5′ and 3′ ends were removed to perform phylogenetic analysis on the common part among the sequences considered. All the phylogenetic analyses detailed in the Materials and Methods section (based either on genetic distances or characters) using DNA or protein sequences, gave congruent results. Thus, we present here only the dendrogram built by the parsimony method on DNA sequences after bootstrapping. In Fig. 1, the *L. major* cathepsin-L like sequence (GenBank Accession no. U43706) was used as outgroup. The dendrogram showed that the *L. donovani* species were more

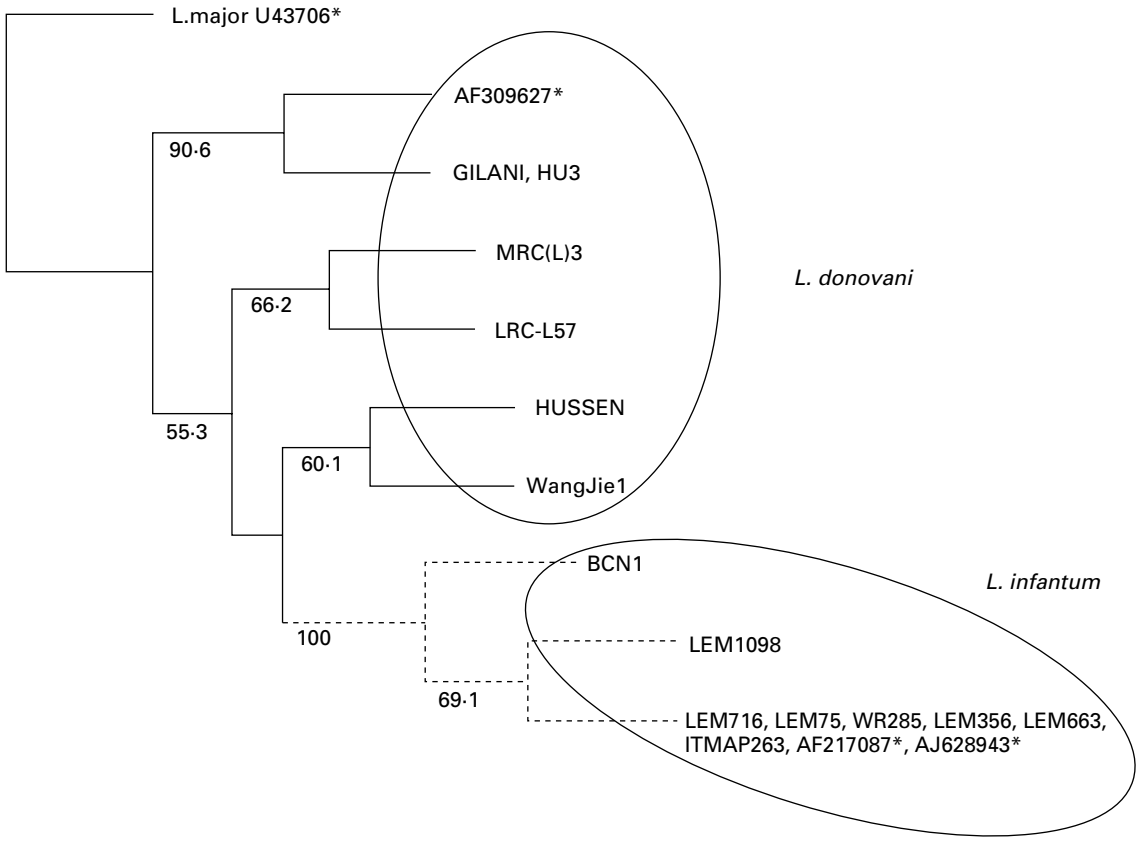

Fig. 1. Phylogenetic analysis of *cpbEF* within the *Leishmania donovani* complex. Dendrogram constructed with *cpbE* and *cpbF* sequences using the parsimony method. Four additional strains were added (*): *L. major* U43706 was used as outgroup and 3 sequences available on GenBank: AF309627 (*L. donovani cpbF*), AF217087 and AJ628943 (*L. infantum cpbE*). Bootstrap values are shown below the branches. Please go to http://www.treedyn.org/hide/hide2005b.html for coloured dendrogram with direct links to GenBank.

polymorphic than *L. infantum*, as illustrated by a genotypic diversity rate of 0·857 (6/7) and 0·3 (3/10), respectively. *L. infantum* strains were clustered together (included AF217087 and AJ628943) with a bootstrap value of 100 and appeared to be a subunit of *L. donovani*. Within the *L. donovani* species, there were 3 groups. The strains GILANI (Sudan), HU3 (Ethiopia) and IS2D (GenBank Accession no. AF309627, Sudan) belonged to the external group (bootstrap value: 90·6). The remainder of the strains were isolated from this group with a weak bootstrap value (55·3) with, on the one hand, 2 Kenyan strains (MRC(L)3 and LRC-L57), and on the other hand the HUSSEN (Ethiopia) and WangJie1 (China) strains. This last group was joined with the *L. infantum* sample (bootstrap value: 33·5). Within *L. infantum*, only 2 of the 4 strains isolated from human cutaneous forms (BCN1 and LEM1098) were separated from the other strains.

Sequence analysis of cpbE *and* cpbF *copies*

Blast analysis using *L. donovani cpbF* (GenBank Accession no. AF309627, 1185 bp) revealed that our *cpbF* PCR product started at nucleotide 411 and finished at 1151. For *L. infantum*, the *cpbE* PCR product went from nucleotides 411 to 1112 of AF217087 and AJ628943 sequences (1146 bp). After translation in protein sequences, the *cpbEF* PCR products did not contain the 136 first amino acids comprising the pre-pro-region and the 12 first amino acids of the mature domain (MD). These PCR products also excluded the 11 last amino acids pertaining to the COOH-terminal extension (CTE). Consequently, after translation, our PCR products corresponded to a protein sequence of 247 aa for CPBF and 234 aa for CPBE, both comprised between the 13th amino acid (aa) of the mature domain (MD) and the 41st aa of the CTE. This PCR did not generate amplification for the other *Leishmania* species samples, representative of the genetic diversity of the genus, and for *Trypanosoma cruzi* and *T. brucei* (unpublished data).

Considering the entire sample, we identified 1 gap and 40 mutation sites among the 16 *cpbEF* sequences, 10 of which led to synonymous mutations (Fig. 2). *CpbE* and *cpbF* differed by a deletion of 39 bp in the mature domain, absent in *cpbE* and by 2 mutations (in the black box on Fig. 1): 1 synonymous mutation in the MD and 1 mutation in position 245 of CTE (nucleotide 699 on Fig. 2), which generated a Proline (CCA) for *L. infantum* and a Leucine (CTA) for

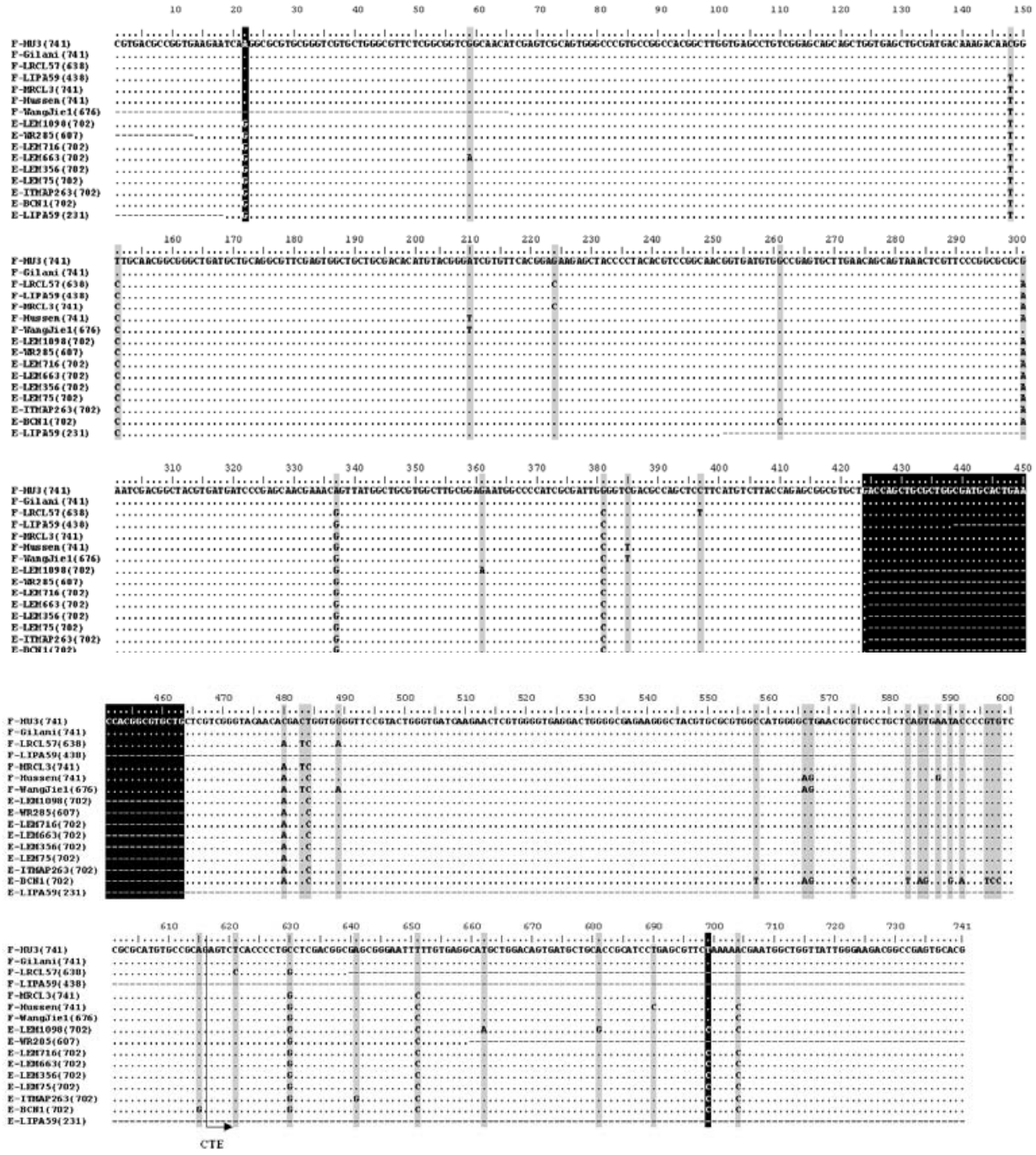

Fig. 2. Multiple alignment of *cpbEF* sequences for the 15 strains pertaining to the *Leishmania donovani* complex. For *L. infantum* strain LIPA59, both *cpbE* and *cpbF* were sequenced. The 40 mutation sites are indicated in grey-shaded boxes except for the 39-bp gap and the 2 mutations that discriminate between *cpbE* and *cpbF*, indicated in black-shaded boxes. Considering the complete *cpbEF* genes, these sequences range from the 411th nucleotide (in the mature domain) to the 1151th for *cpbF* and to the 1112th for *cpbE* (in the COOH-terminal extension). ↳ indicates the 5′ CTE.

L. donovani. Positive selection analysis using the maximum likelihood method (Yang, 1997, 2002) revealed that the second mutation site ($P_{245}L$) was under positive selection (unpublished data). Within the *L. infantum* species, the *cpbE* sequences showed a weak polymorphism, concerning for the most part 2 strains isolated from the cutaneous forms (LEM1098: 3 mutations, BCN1: 14 mutations). Some of these mutations generated amino acids common to *L. mexicana* CPBs (Fig. 3).

Among the 19 mutation sites observed for *L. infantum*, there was only 1 mutation on the LEM663 sequence (synonymous mutation) and 1 on the ITMAP263 sequence (synonymous mutation). Concerning BCN1, 13 out of the 14 mutations were localized at the 3′ end of the mature domain. The

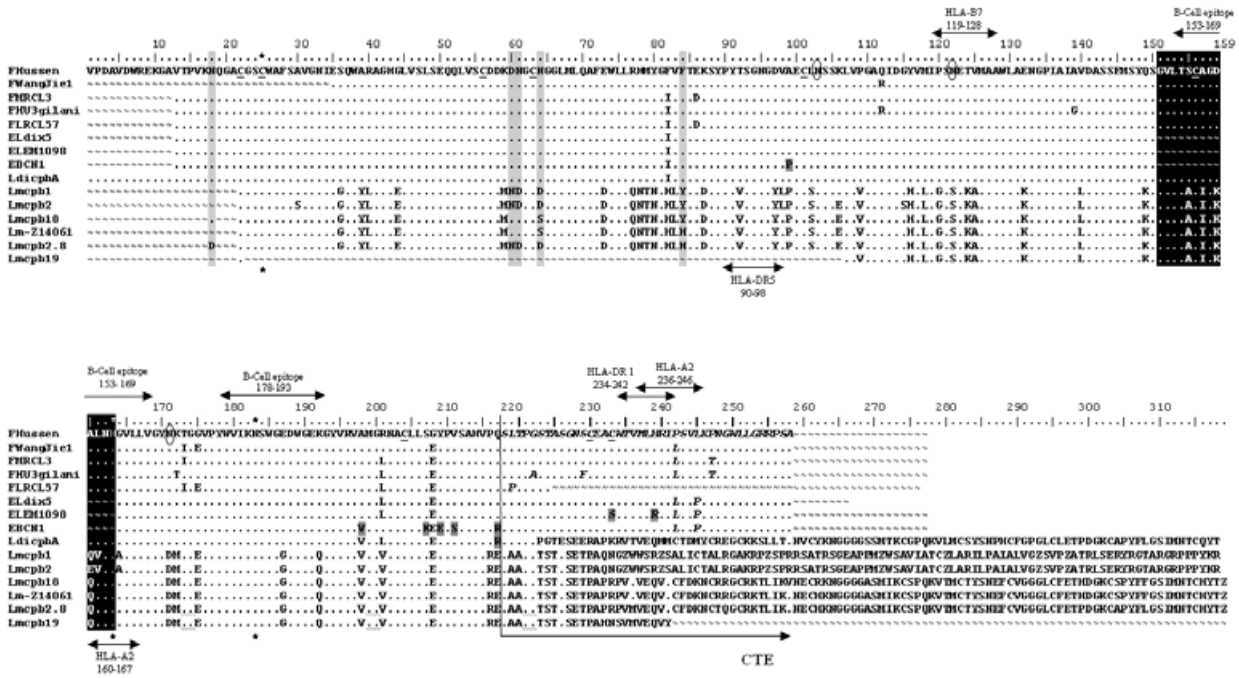

Fig. 3. Comparison of the amino acid sequences of different CPB isoforms: CPBF (referred to as F plus strain name) and CPBE (referred to as E plus strain name) obtained in this study, *L. donovani* CPBA AF309626 (LdicpbA) and 6 *L. mexicana* sequences (LMCPB1, LMCPB2, LMCPB18, LM-Z14061 (LMCPBR), LMCPB2.8, LMCPB19). Dots correspond to identical amino acids. Mature domain and the 41st amino acids of the COOH-terminal extension (in italics) are represented. Cysteine residues are underlined (C), the three potential N-glycosylation sites are surrounded (N) and the 13-aa deletion absent within *L. infantum* CPBE is indicated in the black-shaded box. Residues 18, 60, 61, 64 and 84 are mentioned in grey-shaded boxes. An asterisk (*) indicates the catalytic triad (Cys_{25}-His_{163}-Asn_{183}). Potential T-cell/B-cell epitopes and their position on the amino acid sequence are indicated with arrows. Mutations on the 2 dermotropic strains (BCN1 and LEM1098) are mentioned in dark grey-shaded boxes.

5 other *L. infantum cpbE* sequences were identical (LIPA59, LEM716, LEM75, WR285, LEM356). The *L. donovani* species appeared more polymorphic with 21 mutation sites. Only 2 strains had the same *cpbF* sequence (HU3 and GILANI) and were identical to *L. donovani cpbF* AF309627. As the *L. major* genome has been completely sequenced, a blast analysis revealed that it did not contain *cpbEF* copies. After translation, the 39-bp fragment of *cpbF* corresponded to the amino acid sequence GVLTS-CAGDALNH. It was identified for all the *cpb* sequences of the trypanosomatid family available on GenBank as *L. donovani cpbA* (another copy of the *cpb* cluster), *L. mexicana cpb1*, *cpb2*, *cpb2.8*, *cpb18*, *cpb19*, *L. pifanoi lpcys2*, *T. cruzi* cruzipain and *L. major cpb*-like (data not shown). On the other hand, a blast analysis on CTE revealed that the *cpbEF* CTE (156 nucleotides) had no homology with the other CTE regions (Fig. 3).

Predicted protein analysis

The sequences of *cpbEF* obtained were used to predict mature proteins except for the 12 first amino acids, which were absent from our sequences. We compared the different CPB isoforms within the *L. donovani* complex and also with *L. major* CPB, *L. mexicana* CPB and *T. cruzi* cruzain.

Residue composition

Considering modelling analysis, the CTE was not shown on the model, as the structure of this domain has not been solved. Three-dimensional structures of CPBF (Fig. 4A) and CPBE (Fig. 4B) had been predicted by Comparative Protein Modelling plus CPBE BCN1 (Fig. 4C) to visualize their numerous mutations. Mutations on the BCN1 strain seemed to have an impact on the conformation but no relevant differences were found (Fig. 4C). The location of cysteine residues (Cys) on the protein surface was different between CPBE and CPBF, perhaps because of the deleted sequence. Within the mature domain, there were 7 Cys for the *L. donovani* CPBF and 6 for the *L. infantum* CPBE because Cys_{156} belongs to the deletion of 13 aa (Fig. 3). For each species, there were also 2 Cys residues specific to the *L. donovani* complex in the CTE. CPBF contained 3 disulfide bonds, like *T. cruzi*, in positions Cys_{22}-Cys_{63}, Cys_{56}-Cys_{101} and Cys_{156}-Cys_{204} (Fig. 4A). This last bond was absent in CPBE because Cys_{156} pertained to the deleted sequence GVLTS**C**AGDALNH (Fig. 4B

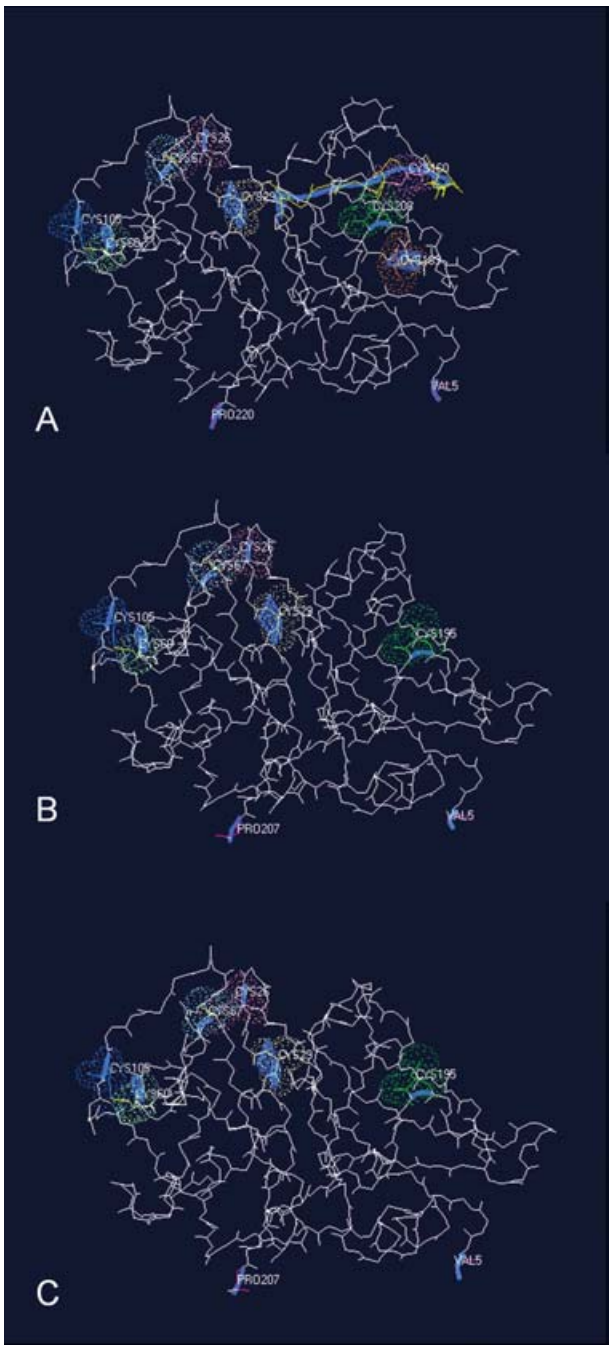

Fig. 4. Homology-based protein model of the mature domain (MD) of *Leishmania donovani cpbF* (A), *L. infantum cpbE* (B) and *L. infantum* BCN1 *cpbE* from cutaneous lesion (C). The blue chain (containing a histidine residue of the catalytic site triad) on (A) corresponds to the 13-aa deletion of *L. infantum*. The modelling server has numbered the Val_1 as Val_5; as a consequence, the residue numbers have a difference of 4 residues compared to those in the text. 5′ MD (Val_5) and 3′-end MD (Pro_{207}-Pro_{220}) are indicated like cysteine residues. Note that *L. infantum* (B and C) does not contain Cys_{160} (pink cloud on A) and Cys_{183} (orange cloud on A). There are 3 disulphide bonds for A (Cys_{160}-Cys_{208} (pink-green), Cys_{60}-Cys_{105} (green-dark blue), Cys_{26}-Cys_{67} (pink-blue) and only 2 for B and C (Cys_{60}-Cys_{105}, Cys_{26}-Cys_{67}).

and C). This sequence was located on the protein surface near the catalytic domain and appeared similar between cruzain and CPBF. There was 100% identity between *L. donovani cpbA* AF309626 and *cpbF* considering protein sequences or DNA of this deleted sequence. Moreover, this sequence contained a histidine residue (His_{163}) belonging to the catalytic triad involved in the protease activity for *T. cruzi* cruzain (McGrath *et al*. 1995) and *L. mexicana* (Juliano *et al*. 2004). This triad (Cys_{25}-His_{163}-Asn_{183}) was found in CPBF, CPBA and for the various CPB isoforms of *L. mexicana* (CPB1, CPB2, CPB2.8, CPB18, CPB19) but not in CPBE, which was missing His_{163} (Fig. 3).

A few amino acid variations between CPB isoenzymes are important in modifying the substrate specificities (Juliano *et al*. 2004; Judice *et al*. 2005). In Fig. 3, we compared these amino acids among, on the one hand, CPBF, CPBA and CPBE for the *donovani* complex and, on the other hand, CPB1, CPB2, CPB2.8, CPB18 and CPB19 for *L. mexicana*. The residues in positions 18, 60 and 61 were identical between *L. donovani* CPBs (A, E and F) and *L. mexicana* CPB18 (i.e. Asn_{18}, Asp_{60} and Asn_{61}) but different for all the other *L. mexicana* CPBs, whereas Asn_{64} and Phe_{84} were specific to the *L. donovani* complex CPBs.

N-glycosylation sites and potential T-cell/B-cell epitopes

CPBE and CPBF contained 3 potential N-glycosylation sites in the mature domain. The first, Asn_{103}, also exists for *Lpcys2* (Boukai *et al*. 2000*a*) and for the different *L. mexicana* CPBs (Fig. 3), whereas Asn_{122} and Asn_{171} were found only on protein sequences of the *L. donovani* complex (Fig. 3). In addition, after blast analysis on the complete *L. major* genome, 8 *cpb*-like copies were revealed and all contained a potential N-glycosylation site in position Asn_{171} (data not shown). Thus, only Asn_{122} appears to be specific to the *L. donovani* complex CPBs but not to CPBE/CPBF since it is present for *L. donovani* CPBA (Fig. 3).

Now considering the complete protein sequence, these predicted CPB isoforms appeared fairly rich in potential T-cell and B-cell epitopes (Fig. 3). Indeed, 2 regions that could provide 3 HLA class I and II epitopes were specific of the *L. donovani* complex: the HLA-B7 epitope (in position 119–128 of the MD), the HLA-A2 epitope (in position 236–246 of the CTE) and the HLA-DR1 epitope (in position 234–342 of the CTE). The HLA-B7 epitope was present in CPBA but HLA-A2 and HLA-DR1 were specific to the CPBE/CPBF CTE. CPBs of the *L. donovani* complex were also characterized by the absence of a few epitopes. In fact, *L. mexicana* CPBs contained a potential HLA-DR5 (in position 90–98 of the MD) and a HLA-A2 (in position 90–98), both

of which are absent for *L. donovani* CPBs (Fig. 3) but also present for *L. major* CPBs-like as well as *L. pifanoi* LPCYS2 (data not shown). Concerning potential B-cell epitopes, 1 region (in position 178–193 of the MD) contained a potential epitope specific to the *L. donovani* complex (CPBA, CPBE and CPBF) and another (in position 153–169) was present for all the CPBs analysed (for *L. mexicana*, *L. major* and *L. mexicana*) except CPBE, because this region contained the 13-aa deleted sequence.

DISCUSSION

Phylogenetic analysis

This study presents an analysis of *cpbE*/*cpbF* sequences for the *L. donovani* complex. The PCR products were sequenced and revealed that the *cpbE* product is specific to *L. infantum*, whereas *cpbF* is specific to *L. donovani*. In the publication by Mundodi *et al*. (2002), the authors revealed the presence of *cpbE* for one *L. donovani* strain, whereas this copy was not evidenced in our study (Mundodi *et al*. 2002). Even so, our PCR is able to reveal both *cpbE* and *cpbF* in the same reaction since a mixed pattern was obtained for *L. infantum* LIPA59, which contains the 2 copies. The phylogenetic tree constructed on the basis of gene sequences follows the species classification since *L. infantum* is individualized from *L. donovani*. We can also note that the taxonomic status of the GILANI (MON30) strain as *L. donovani*, initially typed as *L. infantum* by MLEE, is confirmed by this analysis (Oskam *et al*. 1998; Quispe Tintaya *et al*. 2004; Zemanova *et al*. 2004). Nevertheless, the *L. infantum* sample shows a low level of polymorphism and is included in the *L. donovani* sample. The polymorphic character of *cpbF*, the weak polymorphism within *L. infantum cpbE*, and the phylogenetic classification of the latter as a subunit of the *L. donovani* species confirm the ancestral status of *L. donovani* species already proposed by several authors (Ibrahim and Barker, 2001; Pratlong *et al*. 2001). This suggests that *L. infantum* has recently descended from a *L. donovani* clone and it may have evolved and adapted to the canine reservoir (Ibrahim, 2002). The differences between *cpbE* and *cpbF*, as, for example, the deletion in *cpbE*, appear to reflect this event. Thus, these differences could be directly or indirectly related to the zoonotic character of *L. infantum* and the anthroponotic character of *L. donovani*. Concerning the clinical polymorphism, even if cutaneous leishmaniasis most probably has a multi-factorial origin – a combination of environmental, parasite, host and vector factors – it is interesting to note that the polymorphism within *L. infantum* CPBE is only caused by mutations on strains isolated from the cutaneous form. Furthermore, most of these 'cutaneous' mutations led to amino acids identical to those of *L. mexicana* CPBs, a dermotropic species. Nevertheless, among the 5 *L. infantum* dermotropic strains, 3 of them revealed CPBE sequences similar to those of other *L. infantum* strains (isolated from visceral form). If we consider that the parasite could play a role in expression of clinical signs, these results lead us to several hypotheses. First, the original strain of *L. infantum* could have been a 'dermotropic strain' which consequently might evolve for a much longer time than the 'viscerotropic strain'. This hypothesis is congruent with the presence of common amino acids between *L. infantum* dermotropic strains (BCN1 and LEM1098) and *L. mexicana*. Second, the dermotropic character results from an adaptation of the parasite to its human host. This last proposition is congruent with the diversity observed within dermotropic strains: each mutation may result from a complex interaction between the parasite and its environment, the immune host response and host genetic factors. Another point concerns the presence of both *cpbE* and *cpbF* copies for *L. infantum* LIPA59 isolated from human cutaneous lesion. This strain might result from a hybridization process between the two species *L. infantum* and *L. donovani* resulting from a sexual recombination. Other results based on microsatellite genotyping and sequencing of other *cp* copies (*cpa* and *cpb*) as well as *lpg2*, *amastin* and *iunh*, have confirmed the heterozygous status of this strain LIPA59 (unpublished data).

CPBE and CPBF comparison

CPBs are composed of 3 regions: a pre-pro-domain (from amino acid (aa) 1 to aa 124), a mature domain (from aa 125 to aa 342) and a CTE (variable size according to the CPB isoform). Sequence analysis of these CPBE/CPBF sequences and blast analysis with other trypanosomatids, including other *L. donovani* complex CPBs such as CPBA, revealed that CPBE and CPBF are specific to the *L. donovani* complex. Considering epitopes, our study revealed potential B-cell and T-cell epitopes specific to the *L. donovani* complex as well as potential T-cell epitopes common to different *Leishmania* CPBs (*L. major*, *L. mexicana*, *L. pifanoi*) but absent for the *L. donovani* complex. The implication of these T and B epitopes in host immune response toward leishmaniasis is currently unknown, so further work is necessary to understand the functional consequences of this epitopic diversity. Results obtained on T-cell epitopes are very interesting because cell-mediated immune responses play a role in both protective and counter-protective immune responses (Sacks and Noben-Trauth, 2002).

Concerning the protein conformation, mutations observed in the anthropozoonotic agent *L. infantum* generate important consequences on the protein structure. CPBE presents a deletion in the mature domain affecting essential characteristics existent in other CPBs. Indeed, because of the CPBE deletion,

CPBE misses a potential B-cell epitope in position 153–169. Furthermore, the protease activity of all papain-like proteases is associated with the catalytic triad consisting of a nucleophilic cysteine, a histidine and an asparagine (for example Cys_{25}-His_{163}-Asn_{183} for *Leishmania*) (Selzer *et al*. 1997; Juliano *et al*. 2004). Our study has revealed that, because of CPBE deletion, His_{163} is absent, as is 1 disulfide bond (absence of Cys_{156}). A similar phenomenon was observed in *T. cruzi* where the absence of Cys_{318} leads to the loss of a disulfide bond (Cazzulo *et al*. 1992). Furthermore, these modifications are close to the catalytic site and it could be suggested that CPBE could have different substrate preferences from those of CPBF or cruzain. Thus, thanks to their differences, these 2 isoforms, CPBE and CPBF, may interact differently with the host and consequently reflect the difference in host specificity between *L. infantum* and *L. donovani*. These observations are in agreement with the notion developed by Mottram *et al*. (1997) who suggest that individual CPB isoforms have a distinct interaction with the host. Within the *L. donovani* complex, at least 5 copies of *cpb* seem to exist (Mundodi *et al*. 2002), and independent analysis of enzymatic activities, cellular localization and regulation of each isoform must be conducted in order to enhance our knowledge of these proteins.

Comparison of CPBE/CPBF with other Leishmania *CPBs*

We also compared our results with CPBs of other *Leishmania* complexes, most particularly with *L. mexicana* CPBs, which are the most frequently studied. CPBs are known to be glycosylated at various sites (Parodi *et al*. 1995) and the comparison with other CPB isoforms shows that CPBE/Fs are different: first by the presence of specific potential N-glycosylation site (Asn_{122}), second by additional potential cell epitopes and characteristic CTEs. The 3D model's prediction showed accessibility on the protein surface of this Asn_{122}. Consequently, this site could be important for the protein structure. Directed mutagenesis on potential N-glycosylation sites was performed on *L. pifanoi cpb* (*Lpcys2*) and revealed that glycosylation is not involved in the targeting to the lysosome (Boukai *et al*. 2000*a*). They seem instead to play an important role in folding, stabilization or protease protection (Dwek, 1998). These synapomorphic characteristics could be associated with the specific epidemiology of the *L. donovani* complex (visceral tropism, etc) compared to other species complexes as several authors have highlighted the important role of CPs in pathogenic processes (Alexander *et al*. 1998; Mottram *et al*. 1998; Mahmoudzadeh-Niknam and McKerrow, 2004). Juliano *et al*. (2004) have demonstrated that some amino acids (60, 61, 64 and 84) are important for substrate specificities for *L. mexicana*. Comparison of CPBE/CPBF with *L. mexicana* regarding residues 18, 60 and 61 revealed that CPBE/CPBFs were closer to CPB18. Unfortunately, little is known about CPB18 specificities except that it is expressed in the intracellular amastigote stage and encodes a 47·9-kDa protein that has a high degree of sequence homology with CPB2.8 (Mottram *et al*. 1997). By contrast, residues 64 and 84 are specific to *L. donovani* complex CPBs. Juliano and coworkers (2004) revealed that residues 60, 61 and 64 are located at α-helices that form the wall of the active site cleft and that changes in these residues would modify the electrostatic environment and thus possibly CPB enzymatic activities. Consequences of these amino acid changes on substrate specificity, enzyme activity, etc. require further study.

The CPBE/CPBF COOH-terminal extension

Cpb of trypanosomatids are characterized by the presence of an unusual CTE, absent in mammal cysteine protease (Boukai *et al*. 2000*b*). We also showed that CPBE/CPBFs have a specific CTE, different from the other CPB CTEs of *Leishmania* and that the CPBE/CPBF CTE contains potential epitope sites. The presence of these particular epitopes is an indicator of the involvement of this region in host–parasite interactions, especially with the immune system. This is in agreement with the data obtained by several authors. Chang and McGwire (2002) and Nakhaee *et al*. (2004) hypothesized that *cpbEF* CTE could act as pathoantigen and thus might be involved in some clinical manifestations of visceral leishmaniasis. Nakhaee *et al*. (2004) have shown the importance of *L. infantum cpb* CTE as an immune response target in canine leishmaniasis (Nakhaee *et al*. 2004).

Finally, cysteine protease genes are very complex and significant differences exist depending on the *Leishmania* species. Mottram *et al*. (1997) suggested that the individual CPB isoforms have distinct roles in the parasite's interaction with its host. Consequently, it is important now to focus our work by studying each *cpb* copy independently, also considering the development stage and the parasite species involved. Furthermore, concerning CPBE/CPBF, the synapomorphic characteristics such as specific CTE suggest that these proteins must be studied further in order to understand their potential involvement in visceral leishmaniasis.

M.H. is sponsored by the CNRS, R.B.G. and A.L.B. by the IRD. We are grateful to D. Sereno for a critical reading of this article and to F. Chevenet for his help with Treedyn software. We gratefully thank I. Mauricio, G. Schonian, K. Soteriadou and F. Pratlong, who provided us with some *Leishmania* strains or DNA. The English version of this manuscript was revised by L. Northrup.

REFERENCES

Alexander, J., Coombs, G. H. and Mottram, J. C. (1998). *Leishmania mexicana* cysteine proteinase-deficient mutants have attenuated virulence for mice and potentiate a Th1 response. *Journal of Immunology* **161**, 6794–6801.

Beyrodt, C. G., Pinto, A. R., Freymuller, E. and Barbieri, C. L. (1997). Characterization of an antigen from *Leishmania amazonensis* amastigotes able to elicit protective responses in a murine model. *Infection and Immunity* **65**, 2052–2059.

Boukai, L. K., Da Costa-Pinto, D., Soares, M. J., McMahon-Pratt, D. and Traub-Cseko, Y. M. (2000*a*). Trafficking of cysteine proteinase to *Leishmania* lysosomes: lack of involvement of glycosylation. *Molecular and Biochemical Parasitology* **107**, 321–325.

Boukai, L. K., McMahon-Pratt, D. and Traub-Cseko, Y. M. (2000*b*). Evidence for a recent mutation giving rise to a truncated copy of a cysteine proteinase gene in *Leishmania pifanoi*. *Parasitology International* **49**, 301–307.

Cazzulo, J. J., Martinez, J., Parodi, A. J., Wernstedt, C. and Hellman, U. (1992). On the post-translational modifications at the C-terminal domain of the major cysteine proteinase (cruzipain) from *Trypanosoma cruzi*. *FEMS Microbiology Letters* **79**, 411–416.

Chang, K. P. and McGwire, B. S. (2002). Molecular determinants and regulation of *Leishmania* virulence. *Kinetoplastid Biology and Disease* **1**, 1.

Chevenet, F., Brun, C., Bañuls, A. L., Jacq, B. and Christen, R. (2006). TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7,** 439.

Coombs, G. H. and Mottram, J. C. (1997). Parasite proteinases and amino acid metabolism: possibilities for chemotherapeutic exploitation. *Parasitology* **114**, S61–S80.

Dwek, R. A. (1998). Biological importance of glycosylation. *Developments in Biological Standardization* **96**, 43–47.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Frame, M. J., Mottram, J. C. and Coombs, G. H. (2000). Analysis of the roles of cysteine proteinases of *Leishmania mexicana* in the host-parasite interaction. *Parasitology* **121**, 367–377.

Ibrahim, M. E. (2002). The epidemiology of visceral leishmaniasis in east Africa: hints and molecular revelations. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **96** (Suppl. 1), S25–S29.

Ibrahim, M. E. and Barker, D. C. (2001). The origin and evolution of the *Leishmania donovani* complex as inferred from a mitochondrial cytochrome oxidase II gene sequence. *Infection, Genetics and Evolution* **1**, 61–68.

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. *Molecular and Biochemical Parasitology* **144**, 36–43.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. Munro, H. N.), pp. 21–132. Academic Press, New York.

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. *European Journal of Biochemistry* **271**, 3704–3714.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120.

Mahmoudzadeh-Niknam, H. and McKerrow, J. H. (2004). *Leishmania tropica*: cysteine proteases are essential for growth and pathogenicity. *Experimental Parasitology* **106**, 158–163.

McGrath, M. E., Eakin, A. E., Engel, J. C., McKerrow, J. H., Craik, C. S. and Fletterick, R. J. (1995). The crystal structure of cruzain: a therapeutic target for Chagas' disease. *Journal of Molecular Biology* **247**, 251–259.

McKerrow, J. H., Engel, J. C. and Caffrey, C. R. (1999). Cysteine protease inhibitors as chemotherapy for parasitic infections. *Bioorganic and Medicinal Chemistry* **7**, 639–644.

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

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. *The Journal of Biological Chemistry* **272**, 14285–14293.

Mottram, J. C., Souza, A. E., Hutchison, J. E., Carter, R., Frame, M. J. and Coombs, G. H. (1996). Evidence from disruption of the *lmcpb* gene array of *Leishmania mexicana* that cysteine proteinases are virulence factors. *Proceedings of the National Academy of Sciences, USA* **93**, 6008–6013.

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

Nakhaee, A., Taheri, T., Taghikhani, M., Mohebali, M., Salmanian, A. H., Fasel, N. and Rafati, S. (2004). Humoral and cellular immune responses against type I cysteine proteinase of *Leishmania infantum* are higher in asymptomatic than symptomatic dogs selected from a naturally infected population. *Veterinary Parasitology* **119**, 107–123.

Oskam, L., Pratlong, F., Zijlstra, E. E., Kroon, C. C., Dedet, I. P., Kager, P. A., Schonian, G., Ghalib, H. W., El-Hassan, A. M. and Meredith, S. E. (1998). Biochemical and molecular characterization of *Leishmania* parasites isolated from an endemic focus in eastern Sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92**, 120–122.

Parodi, A. J., Labriola, C. and Cazzulo, J. J. (1995). The presence of complex-type oligosaccharides at the C-terminal domain glycosylation site of some molecules

of cruzipain. *Molecular and Biochemical Parasitology* **69**, 247–255.

Pratlong, F., Dereure, J., Bucheton, B., El-Saf, S., Dessein, A., Lanotte, G. and Dedet, J. P. (2001). Sudan: the possible original focus of visceral leishmaniasis. *Parasitology* **122**, 599–605.

Quispe Tintaya, K. W., Ying, X., Dedet, J. P., Rijal, S., De Bolle, X. and Dujardin, J. C. (2004). Antigen genes for molecular epidemiology of leishmaniasis: polymorphism of cysteine proteinase B and surface metalloprotease glycoprotein 63 in the *Leishmania donovani* complex. *Journal of Infectious Diseases* **189**, 1035–1043.

Sacks, D. and Noben-Trauth, N. (2002). The immunology of susceptibility and resistance to *Leishmania major* in mice. *Nature Reviews Immunology* **2**, 845–858.

Sajid, M. and McKerrow, J. H. (2002). Cysteine proteases of parasitic organisms. *Molecular and Biochemical Parasitology* **120**, 1–21.

Selzer, P. M., Chen, X., Chan, V. J., Cheng, M., Kenyon, G. L., Kuntz, I. D., Sakanari, J. A., Cohen, F. E. and McKerrow, J. H. (1997). *Leishmania major*: molecular modeling of cysteine proteases and prediction of new nonpeptide inhibitors. *Experimental Parasitology* **87**, 212–221.

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 Letters* **311**, 124–127.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**, 4876–4882.

World Health Organization (2002). Annex 3: burden of disease in DALYs by cause, sex and mortality stratum in WHO regions, estimates for 2001, pp. 192–197. World Health Organization, Geneva.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555–556.

Yang, Z. (2002). Inference of selection from multiple species alignments. *Current Opinion in Genetics and Development* **12**, 688–694.

Zemanova, E., Jirku, M., Mauricio, I. L., Miles, M. A. and Lukes, J. (2004). Genetic polymorphism within the *Leishmania donovani* complex: correlation with geographic origin. *The American Society of Tropical Medicine and Hygiene* **70**, 613–617.