J Mol Model (2013) 19:3053–3064
DOI 10.1007/s00894-013-1834-8

ORIGINAL PAPER

Gene identification and comparative molecular modeling of a *Trypanosoma rangeli* major surface protease

Paulo H. M. Calixto · Mainá Bitar · Keila A. M. Ferreira · Odonírio Abrahão Jr. · Eliane Lages-Silva · Glória R. Franco · Luis E. Ramírez · André L. Pedrosa

Received: 26 November 2012 / Accepted: 20 March 2013 / Published online: 13 April 2013

Abstract *Trypanosoma rangeli* is a hemoflagellate parasite which is able to infect humans. Distinct from *Trypanosoma cruzi*, the causative agent of Chagas disease, *T. rangeli* is non-pathogenic to the vertebrate host. The manner by which the *T. rangeli* interacts with the host is still unknown, but it certainly depends on the surface molecules. Major surface proteins (MSP) are GPI-anchored, zinc-dependent metalloproteases present in the surface of all trypanosomatids studied so far, which are implicated as virulence factors in pathogenic trypanosomatids, such as *Leishmania* spp and *T. cruzi*. The aims of this work were to generate the complete sequence of a *T. rangeli* MSP (TrMSP) gene and to determine the 3D-structure of the predicted protein by homology modeling. The plasmid bearing a complete copy of a TrMSP gene was completely sequenced and the predicted protein was modeled using Modeller software. Results indicate that TrMSP open reading frame (ORF) codes for a predicted 588 amino acid protein and shows all elements required for its posttranslational processing. Multiple sequence alignment of TrMSP with other trypanosomatids' MSPs showed an extensive conservation of the N-terminal and central regions and a more divergent C-terminal region. *Leishmania major* MSP (LmMSP), which had its crystal structure previously determined, has an overall 35 % identity with TrMSP. This identity allowed the comparative molecular modeling of TrMSP, which demonstrated a high degree of structural conservation between MSPs from other trypanosomatids (TrypMSPs). All modeled MSPs have a conserved folding pattern, apart from structural divergences in the C-domain and discrete differences of charge and topology in the catalytic cleft, and present the same geometry of the canonical HE*XX*H zinc-binding motif. The determination of surface charges of the molecules revealed that TrMSP is a predominantly positive protein, whereas LmMSP and *Trypanosoma cruzi* MSP (TcMSP) are negative proteins, suggesting that substrates recognized by TcMSP and LmMSP could not interact with TrMSP. Moreover, the comparison between TrMSP and TcMSP protein sequences has revealed 45 non-neutral amino acid substitutions, which can be further assessed through protein engineering. The characteristics of TrMSP could explain, at least in part, the lack of pathogenicity of *T. rangeli* to humans and point to the necessity of identifying the biological targets of this enzyme.

Keywords Comparative modeling · Host-parasite interaction · Major surface protease (MSP or gp63) · Non-neutral substitutions · *Trypanosoma rangeli*

Electronic supplementary material The online version of this article (doi:10.1007/s00894-013-1834-8) contains supplementary material, which is available to authorized users.

P. H. M. Calixto · K. A. M. Ferreira · O. Abrahão Jr. · A. L. Pedrosa (✉)
Departamento de Bioquímica, Farmacologia e Fisiologia, Instituto de Ciências Biológicas e Naturais, Universidade Federal do Triângulo Mineiro, Uberaba, Brazil
e-mail: pedrosa@icbn.uftm.edu.br

M. Bitar · G. R. Franco
Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

E. Lages-Silva · L. E. Ramírez
Departamento de Imunologia, Microbiologia e Parasitologia, Instituto de Ciências Biológicas e Naturais, Universidade Federal do Triângulo Mineiro, Uberaba, Brazil

Introduction

Trypanosoma cruzi is a flagellate protozoan that infects humans and other mammals and is the etiologic agent of Chagas disease in Central and South America. *T. cruzi* is transmitted by Triatomina of the genera *Rhodnius* and *Triatoma* or, more rarely, by vertical transmission and transplantation of organs and/or tissues [1]. *Trypanosoma rangeli* is sympatric with *T. cruzi* and transmitted by the same insect vectors, however, it is harmless to the vertebrate host. Nevertheless, *T. rangeli* has great medical and epidemiological importance, since it shares several antigenic determinants with *T. cruzi*, culminating in serum cross-reactivity in the diagnosis of Chagas disease [2].

The surface proteins of trypanosomatids are fundamental molecules for the host/parasite interaction. One of these molecules is the major surface protease (MSP), a zinc-dependent metalloprotease, initially described in the genus *Leishmania* as a 63 kDa glycoprotein (gp63) that can either be associated with the membrane via a glycosylphosphastidylinositol anchor (GPI) [3] or secreted, without the attachment signal for the GPI anchor [4]. The MSP catalytic site is composed by a HE*XX*H motif coordinated by a Zn^{2+} ion [5]. These proteases have been implicated as virulence factors of *Leishmania* spp., given its interaction with a large variety of substrates. The functions of MSPs include: i) complement system inactivation, through the cleavage of C3b in its inactive form, C3bi [6]; ii) cleavage of the CD4 receptor of T lymphocytes interfering with cell-mediated immunity [7]; iii) interaction with macrophage receptors, ensuring the parasite entrance without respiratory burst [8]; and iv) cleavage of mTOR (mammalian target of rapamycin), a major translational control element, generating a state of low responsiveness of the infected macrophage, facilitating intracellular replication of *Leishmania* [9].

Interestingly, the MSPs have been described in other trypanosomatids such as *Trypanosoma brucei* and *T. cruzi*. In *T. brucei*, MSPs are responsible for the cleavage and release of variant surface glycoproteins (VSG) in the antigenic switching mechanism [10]. In *T. cruzi*, MSPs are encoded by a gene family with approximately 251 copies [11]. The molecular bases involved in the action of MSPs in *T. cruzi* and its function are still unknown. However, neutralization experiments with anti-MSP antibodies decreased the infectivity of *T. cruzi* in VERO cells by 50 %, suggesting that MSPs are a virulence factor in this specie [12].

Although MSPs are encoded by gene families in all sequenced trypanosomatids, there is only one solved crystal structure: the MSP of *L. major* (PDB code: 1LML) [5]. As an alternative to crystallographic studies, protein structures can be solved *in silico*, by comparative modeling. In some cases, *in silico* modeling allows the generation of tridimensional structure of proteins with resolution comparable to experimental results. The comparative modeling technique predicts the three-dimensional structure of a protein based on amino acid similarities with the sequence of a related protein of known structure. This is considered to be the most accurate computational method for structural characterization of proteins [13].

Although the molecular bases of host/parasite interactions in *T. rangeli* are not known, it is presumable that surface protein families present in other trypanosomatids must have a fundamental role in the related mechanisms. Despite its importance, functional details of most *T. rangeli* surface proteins are still unknown and comprehensive studies on protein structure could help the understanding of processes that determine pathogenicity of trypanosomes in the vertebrate host.

To our knowledge, there are no reports describing genes encoding MSPs in *T. rangeli*. The first evidence of MSPs genes in this species arose from the construction of a genomic library, where 12 gene fragments with a high percentage of similarity with *T. cruzi* MSPs were identified [14]. In this paper, we report the first complete sequence of the gene encoding a *T. rangeli* MSP (TrMSP), obtained from a library of clones through Genome Sequence Survey (GSS). From the predicted protein sequence, we generated and validated the three-dimensional model of TrMSP. Subsequently, we evaluated the distribution of surface charges to better characterize the model and predicted functionally important residues that can be a target for further protein engineering studies.

Materials and methods

Sequencing of a complete MSP gene in *T. rangeli*

The nucleotide sequence of *T. rangeli* (P07 strain) MSP gene was obtained in a recently published library [14]. Bacterial clones containing plasmids with the forementioned sequence and resistant to ampicilin were grown in LB medium (20 g L^{-1} tryptose, 5 g L^{-1} yeast extract, 8.5 mM NaCl) in the presence of 50 μg mL^{-1} ampicilin and incubated for 18 h at 37 °C under constant agitation. Cells were ruptured through alcaline lysis and recombinant DNA purification was performed as described elsewhere [15].

The only plasmid (2C01) comprising the complete sequence of the MSP gene was sequenced through primer walking with the following oligonucleotides: 1) Tr2C01-PW1 5′-TGGAACATCGCTGCACGT-3′; 2) Tr2C01-PW2 5′-CAGCACATCTCTGCAGG-3′; 3) Tr2C01-PW3-5′ GATTACTGCCCCATCATCGT-3′. All sequencing reactions were performed using the Big Dye Terminator Cycle Sequencing kit, version 3.1 (ABI PRISM™) and processed with the 3130 XL

Genetic Analyzer (Applied Biosystems), according to manufacturer instruction.

Computational analysis of predict TrMSP sequences

The search of *cis* elements involved in TrMSP processing was performed using the following tools: 1) Signal IP 3.0 Server™ [16] for the identification of signal peptide; 2) PredGPI™ [17] for the prediction of GPI anchor addition site; 3): ProtParam™ [18] for the determination of MSP isoeletric point. The amino acid sequences of different MSPs were aligned using ClustalW and manually curated (http://www.ebi.ac.uk/Tools/msa/clustalw/) [19].

Bioinformatics analysis of TrMSP structural characteristics

The complete *T. rangeli* MSP sequence [GenBank JQ579649] was submitted to BLAST [20] searches against publicly available databases in order to identify possible homologues and determine similarity levels between the MSP of *T. rangeli* and other MSPs, especially those from trypanosomatids. The criteria for detecting homology included *e*-value, alignment scores and sequence coverage. Once the set of MSP homologues in trypanosomatids was established, all subsequent steps were simultaneously performed with all sequences.

Sequence alignments were generated, using Promals3D [21] with default values and manually curated and secondary structure predictions were performed using PSI-PRED [22]. The tertiary structure for TrMSP was not experimentally determined and therefore three-dimensional models were generated through comparative modeling for this protein and for the MSPs of all TrypMSPs. BLAST searches against the PDB database were performed in order to identify a suitable structural template for all MSP proteins. Sequence alignments between the target proteins and the template were generated using Promals 3D and the conservation of amino acids with similar physico-chemical properties and gap insertions were inspected with the support of DNATagger, a visualization tool. Modeller (version 9.10) [23] was used to generate a preliminary set of 100 candidate structures for each MSP. The Zn^{2+} atom and a conserved water residue (corresponding to Zn^{2+} 578 and water 663 in 1LML) were both included in the comparative modeling protocol as well as all nine disulfide bridges present in the template structure (Cys residues are conserved in all sequences). For each of the MSPs, all candidate structures were ranked according to energetic and stereochemical features obtained from Procheck [24] Ramachandran plots and ProSA [25] Z-score values, respectively. We have defined a proportionality score to assess the quality of candidate structures, consisting of a comparison between the stereochemical features analyzed by Procheck through the percentage of residues in most favored regions of the Ramachandran plot (*R*-value) and the energy evaluation represented by the absolute ProSA Z-score value (*P*-value). As the *R*-value should be as high as possible (positive and up to 100) and the *P*-value should be as close to zero as possible (either positive or negative), a reasonable evaluation seems to be the ratio between *R*-value and *P*-value, a score we have named *S*-value. The higher the *S*-value, the better the quality of the candidate structure. From this evaluation, the best-quality model for each of the proteins was defined as being the candidate structure with highest *S*-value and used in subsequent analysis. Subsequently, we performed loops and side-chain modeling. Structures were minimized using OPLS 2005 force field with additional parameters provided by Schrödinger (www.schrodinger.com) in the MacroModel software. Validation results are showed in Supplementary Table 1.

Assignment of TrypMSP surface charges

Surface charges for all surface models were assigned using the PDB2PQR [26] server using the AMBER [27] force field and naming scheme. PROPKA [28] was used for protonation, considering the physiological pH (7.0). Results were submitted to APBS calculations. All programs are available within the Kryptonite server (http://kryptonite.nbcr.net/apppdb2pqr) [29]. Visualization of the charged surfaces were performed in PyMOL [30] using the APBS tools plugin.

Identification of non-neutral amino acid substitutions

A pairwise sequence alignment was built between TrMSP and TcMSP (its closest homologue among TrypMSPs considered in this study) using Promals3D [21]. A Perl script was used to identify all amino acids substitutions among these two proteins and generate a mutation list as input for SNAP [31], a neural-network based algorithm to evaluate the effects of such substitutions to protein function. The minimum reliability index for SNAP was set as 0 and the minimum expected accuracy as 50 %.

Results and discussion

Identification of the TrMSP gene and analysis of the predicted protein

The insert of pUC19-2C01 clone, isolated from the *T. rangeli* genomic library was sequenced through primer walking. From the determined consensus sequence it was possible to identify a 1767 bp ORF that encodes a hypothetical 588aa protein (Fig. 1) with a predicted pI of 7.13. BLAST analysis with blastx revealed 58 % and 35 % identity between the predicted *T. rangeli* protein and MSPs of *T.*

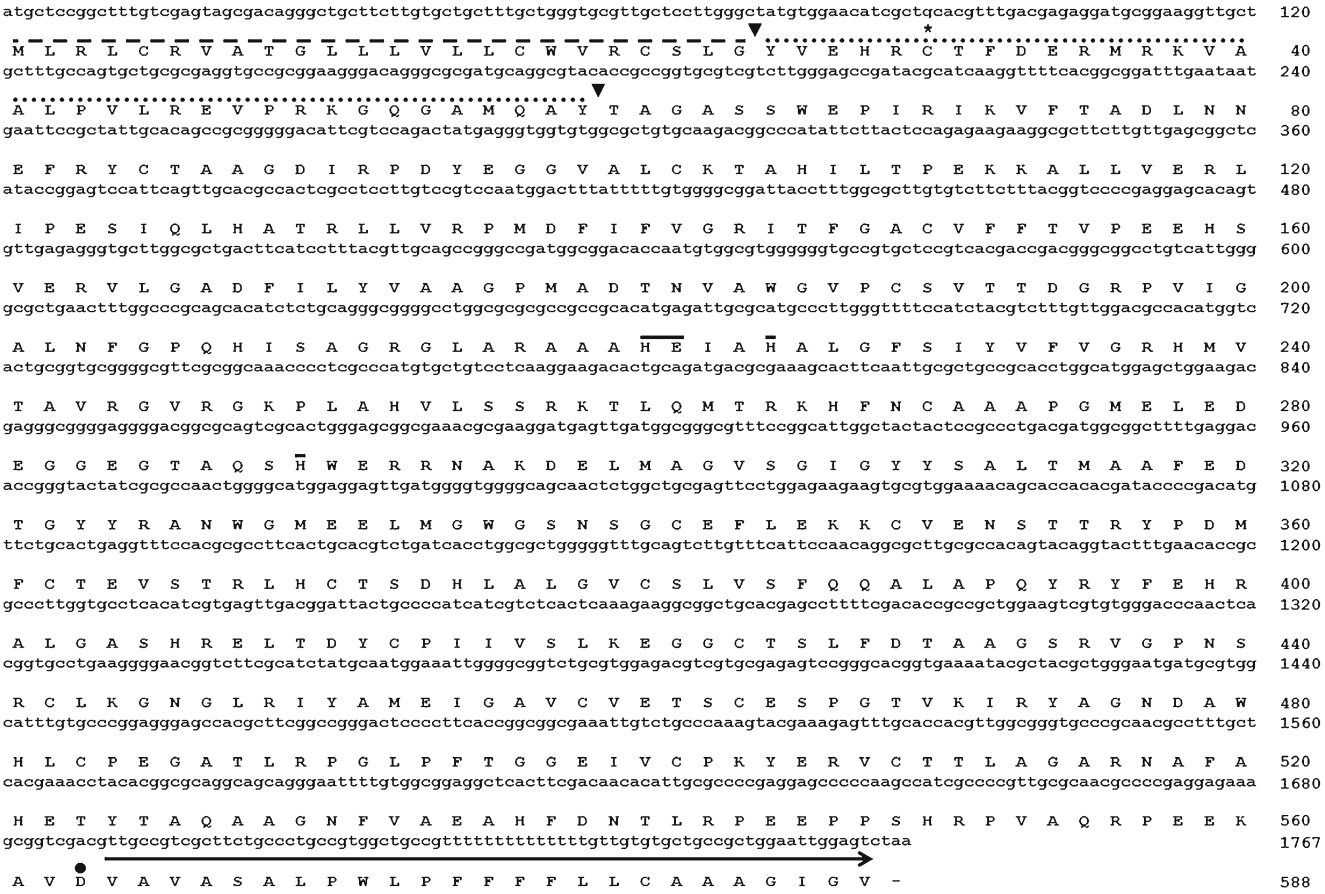

Fig. 1 Nucleotide sequence and the corresponding amino acid sequence of TrMSP protein. Symbols and lines are: *dashed line*—signal peptide; *dotted line* - pro-peptide; *asterisk*—the probable cysteine residue involved in the process of cysteine switch (Cys30); *arrowheads*—probable signal peptide and pro-peptide cleavage sites; *continuous line*—catalytic residues (His221, Glu222, His225, and His290); *arrow*—GPI anchor signal sequence; *bullet sign*—omega site (Asp563) for GPI anchor addition

cruzi (TcMSP) and *L. major* (LmMSP), respectively. The coverage between TrMSP/TcMSP and TrMSP/LmMSP were 92 % and 75 %, respectively. From these results, we assumed that the ORF from clone pUC19-2C01 does in fact correspond to a *T. rangeli* MSP.

MSPs suffer several post-translational modifications [5] and human knowledge regarding protein processing is essential for the proper prediction of tertiary structures of mature proteins. These predictions are especially valuable in cases where there is a low percentage of sequence identity between query and template. Post-translational modifications of TrypMSPs, such as the identification of the signal peptide, the propeptide and the GPI-attachment site, were predicted with the use of several software programs. According to SignalP software prediction [16], residues 1 to 24 of the predicted TrMSP protein correspond to a signal peptide (Fig. 1). PredGPI software [17] indicates residue Asp563 as the omega site for addition of the GPI anchor (Fig. 1). Based on these data, we can infer that the TrMSP protein has all elements required for its localization in the outer surface of *T. rangeli*.

Furthermore, the alignment of the amino acid sequences of TrMSP and TcMSP revealed that the former has a conserved N-terminus extension, the propeptide, which contains a cysteine involved in a switch mechanism that controls enzymatic activity, avoiding the degradation of intracellular components by the newly synthesized enzyme (Fig. 1).

Conservation of MSP sequences

All MSPs studied to date, in different species, showed a high degree of sequence conservation with all other MSPs [32–34]. The only exception is the predicted sequence of *Leishmania guyanensis* MSP that apparently lacks the last five cysteine residues, however, no studies on the viability or functionality of this enzyme have been conducted [35]. In accordance to this observation, multiple sequence alignment of MSPs from *T. rangeli*, *T. cruzi*, *T. brucei* (TbMSP), *L. major*, and *Crithidia fasciculata* (CfMSP) revealed a high amino acid conservation between sequences, including the active residues HE*XX*H and a third histidine residue involved in the formation of the catalytic site (Supplementary Fig. 1). In general, approximately

16 % of all residues are identical in all investigated MSPs, including 18 cysteine, 12 glycine and seven proline residues, which are conserved in all MSPs (Supplementary Fig. 1). The multiple sequence alignment also revealed that TrMSP is closest related to TcMSP, with 58 % identity between amino acid sequences. Within the *Leishmania* genus, TrMSP is closer to the LmMSP than to any other MSP in this genus (Supplementary Fig. 2).

Comparative modeling of TrMSP

To generate a structural homology-based model of TrMSP we selected as template the structure of leishmanolysin (LmMSP, PDB code: 1LML) solved by X-ray diffraction with a resolution of 1.86 Å [5]. We performed several searches of TrypMSPs amino acid sequences in PDB DataBank, using the complete and partial regions of a few identities as queries. However, no structure other than leishmanolysin was retrieved. PDBReport indicated that 1LML has the quality features that enable it to be used as template, with the following global quality results: first generation packing quality of −0.77; second generation packing quality of −0.801; Ramachandran plot appearance of −1.307; chi-1/chi-2 rotamer normality of −0.929; backbone conformation of −0.877. Sequence alignment showed that TrMSP has 35 % identity and 58 % similarity with LmMSP residues. Gaps were closed by comparing the TrypMSP sequences with 1LML via ClustalW and Promals3D, both in the default mode. TrypMSP protein sequences were manually processed by removing of elements lost during post-translational modifications, such as signal peptide, propeptide, and the residues downstream of

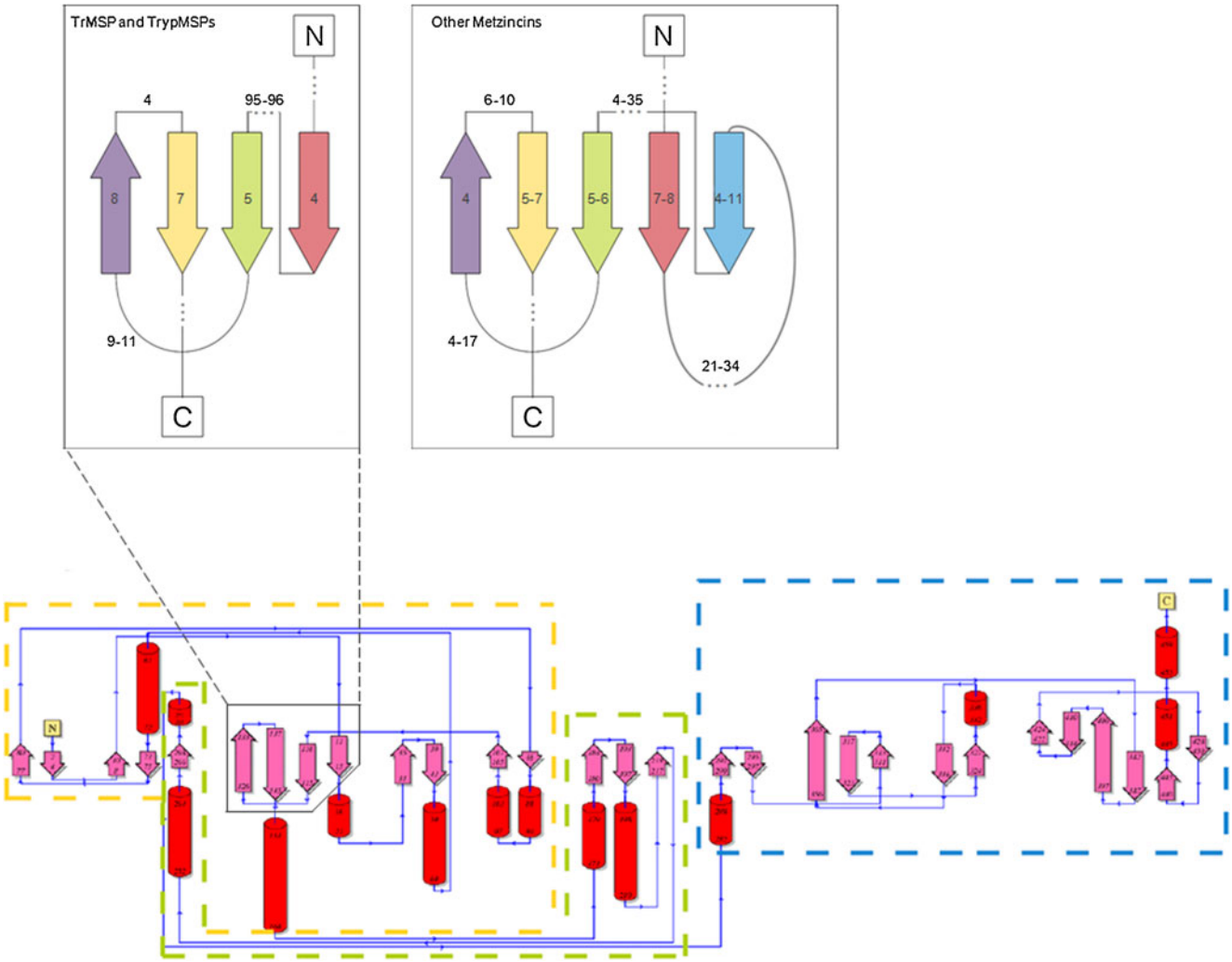

Fig. 2 Topological and structural analyses of TrMSP protein. A) Topology of the TrMSP protein derived from PDBSum, evidencing the secondary structure elements that compose the protein structure. Structural subdomains are delimited inside *dashed lines*, as follows: C-terminal domain in *blue*, central domain in *green* and N-terminal domain in *yellow*. The twisted β-sheet region is enlarged in a schematic representation and compared with the correspondent region of other metzincins, evidencing the absence of the second strand in TrMSP when compared to other proteins from this family. *Numbers* indicate the length range of each strand of segments between strands

the omega-site, which are not represented in the template 1LML because this structure corresponds to the mature protein [5].

Then, alignments were generated for models construction. After the generation of protein structures bearing the zinc ion in the catalytic site and quality evaluation regarding structural energy and stereochemical features, a three-dimensional model for TrMSP was selected and further submitted to molecular dynamics simulations for structural stability assessment. The proposed protein structure is globular and comprised by 29 β-strands and 14 α-helices, according to PDBSum [36] data (accession number c834 and password 210521, Fig. 2, bottom panel).

Zinc-dependent metalloendopeptidases that present the catalytic motif HE*XX*H can be grouped into three main groups: gluzincins, aspzincins and metzincins, which include astacins, serralysins, adamalysins, MMPs, snapalysins and leishmanolysins [37]. All families present a characteristic catalytic domain formed by the two α-helices, named A and B [38], packed against one side of a five-stranded twisted β-sheet [5] (in which the second strand is absent in TrMSP and LmMSP, Fig. 2, top panel). The HE*XX*H motif is placed in the B α-helix and the side chains of both histidines coordinate the zinc atom in the active site. In contrast to the folding pattern of other metzincin family members, TrMSP and LmMSP have only four β-strands in the catalytic module (Fig. 2). The missing β-strand is replaced by an ~40-residue insertion in LmMSP and TrMSP.

The predicted TrMSP structure is compact and can be divided into three main structural subdomains: N-terminal, central and C-terminal (Fig. 3). The N-terminal subdomain corresponds to residues Thr60 to Phe229 of the processed protein and possesses a folding pattern similar to the catalytic module of other zinc-proteases. This subdomain contains the active residues from the catalytic site HE*XX*H (His221, Glu222 and His225), the signature motif of zinc proteases. This subdomain also bears a deep cleft in the interface of the active site, which is located between the N-terminal and central subdomains of TrMSP protein. The HE*XX*H motif is located in an α-helix and the side chains of the two histidines coordinate the zinc ion in the catalytic site (Fig. 3). TrMSP has four β-strands in the catalytic cleft (residues 70 to 74 and 169 to 204), different from other metzincins, which have five β-strands, but similar to leishmanolysin which is the closest protein to TrMSP regarding sequence similarity which has already been classified. It is worth noticing that the antiparallel β-strand (β-strand IV) is still present in TrMSP, given its importance in the catalytic mechanism [37]. The missing β-strand (β-strand II) is replaced by an ~40-residues insertion in both LmMSP and TrMSP. Additionally, the subdomain also bears a long α-helix positioned in parallel to the shorter α-helix containing the active residues.

The central subdomain corresponds to residues Ser230 to Gly341 and has a compact folding pattern with anti-parallel helices and anti-parallel β-strands, forming its core (Fig. 3). A single disulfide bond (Cys270-Cys342) is responsible for linking the central- and carboxyl-subdomains. Members of the metzincin family have the extended motif HE*XX*H*XX*G*XX*H, where the glycine residue is part of a small loop that projects the third histidine residue inside the catalytic site, placing its side chain in coordination with the zinc atom. Interestingly, TrMSP has a 61-residue insertion between the Gly228 of the extended motif and the His290, which also coordinates with the zinc atom (Fig. 3). A similar insertion is present in snapalysins and leishmanolysin, making difficult the definition of the metzincins class based on sequence data alone [37].

The geometry of residues coordinating the zinc atom is similar to that of other zinc proteases. The zinc atom is coordinated by the side chains of His221, His225 e His290 at distances of 2.7 Å, 2.3 Å, and 2.7 Å, respectively. In addition to the three histidine residues in the catalytic site, there is a water molecule in coordination with the zinc atom, at a distance of 2.7 Å (relative to the oxygen atom). This water molecule is located very close to the zinc atom and is the nucleophile responsible for attacking the substrate peptide

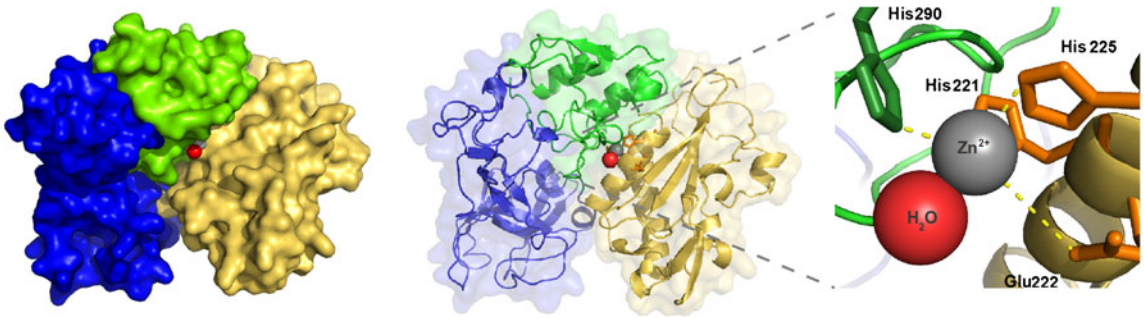

Fig. 3 Protein structure of TrMSP as obtained by comparative modeling. Domains are colored according to the topology in Fig. 2 (C-terminal domain in blue, central domain in green and N-terminal domain in *yellow*). The catalytic site is enlarged for a more detailed representation. The zinc atom and the catalytic water molecule are displayed as *gray* and *red spheres*, respectively. Residues from the catalytic site that coordinate the zinc ion are shown in licorice and colored in *darker shades*

bond. TrMSP active site seems to be pentacoordinated with Glu222, at a distance of 1.901 Å (Fig. 3).

The C-terminal subdomain, corresponds to the region between amino acids Cys342 and Ala518. Within this subdomain are located six out of nine disulfide bonds of the structure which indicates an increased rigidity of this domain. The GPI anchor is added in the last residues of the C-terminal and is characterized as a spiral loop, similar to an α-helix.

Conservation of TrMSP tertiary structure

For comparison, we further investigated the conservation of TrMSP tertiary structure compared with MSPs obtained from other trypanosomatids (TrypMSPs). As the only experimentally solved structure available for comparison was LmMSP (leishmanolysin, 1LML), we determined structural models of MSPs from *C. fasciculata*, *T. brucei* and *T. cruzi* (CfMSP, TbMSP and TcMSP, respectively). All models were constructed and refined by homology-based structural modeling, with LmMSP as template and using the same computational protocol employed for TrMSP. These protein models are publicly available in the Protein Model Database (PMDB) [39], under the ID numbers: PM0078813 for TrMSP, PM0078814 for CfMSP, PM0078815for TbMSP, and PM0078816 for TcMSP.

From the comparison between the obtained structures (Fig. 4), we can observe that the MSPs from trypanosomatids are highly conserved, sharing the N-terminal fold and central domains, including the catalytic site. On the other hand, the C-domain has structural divergences, although the overall folding pattern is conserved. The analysis of TrypMSPs surfaces revealed that the cleft in the catalytic site is slightly different in each protein, regarding charge and topology (Figs. 4 and 5). These differences suggest that each MSP analyzed has specificity for distinct substrates, and this is strongly supported by the fact that MSPs have different functions in each specie [10, 12, 40, 41].

Determination of surface partial charges of TrMSP

Considering that the enzyme/substrate interaction is influenced not only by topological aspects, but also by the complementarity of charges in the contact surface, we have decided to investigate the map of surface partial charges of LmMSP, CfMSP, TbMSP, TcMSP and TrMSP (Fig. 5). Regarding the catalytic site, all TrypMSPs were mainly positive, due to the positivity of the zinc atom and the histidine residues. The overall surface charge, on the other hand, is predominantly negative for LmMSP (−14e), CfMSP (−19e), TbMSP (−20e) and TcMSP (−21e). Surprisingly, the analysis of TrMSP surface charges has revealed substantial differences when compared to all other TrypMSPs. For TrMSP, partial surface charges are more electropositive (+2e), as presented in Fig. 5. This is a remarkable finding, since TrypMSPs have their primary, secondary, and tertiary structures conserved over >200 Ma of evolution and seem to differ solely in the distribution of their surface charges. The negativity of the surface around the catalytic sites of TrypMSP may play two roles: (i) to prevent the binding of incorrect substrates in the catalytic cleft by means of electrostatic repulsion and (ii) to facilitate the interaction of specific substrates with the enzyme, projecting these molecules into the catalytic cleft [42]. Taking these results together, the topology of the catalytic site and, mainly, the differences in the electrostatic potential in the surface of the catalytic cleft, we can suggest that substrates recognized by other TrypMSPs would not interact with TrMSP. This hypothesis is supported by and corroborates with the description of the distinct roles that MSPs play in different species [10, 12, 41, 43].

Identification of important residues in TrMSP

Interestingly, among all TrypMSPs studied herein, TcMSP is the closest homologue to TrMSP. The few differences between these two proteins may account for the lack of pathogenicity of *T. rangeli*. To assess this hypothesis, we have used prediction strategies to identify different residues between the two proteins that may lead to structural (and therefore functional) changes. After a pairwise sequence alignment, we observed 257 amino acids substitutions between TcMSP and TrMSP (Fig. 6). From these substitutions, 45 were classified as non-neutral with an average expected accuracy of 62 %, hypothetically indicating a localized difference in the protein structure and function (Fig. 6). These non-neutral substitutions are located along the TrMSP protein sequence apparently randomly, although some regions have a higher concentration of such residues. From the 45 non-neutral amino acids substitutions identified, 11 (~25 %) are of negative residues in TcMSP to non-negative (either polar not charged, apolar or positively charged) residues in TrMSP (and around 10 % of all 257 substitutions are of this type) and nine (20 %) are substitutions of non-positive residues TcMSP to positive residues in TrMSP (and around 7 % of all 257 substitutions are of this type). From all 257 substitutions, more than 25 % (67 positions) are involved in the charge change (from highly negative to slightly positive) of TrMSP when compared to TcMSP. Especially, there is an increase in the number of arginines and a drastic decrease in the number of aspartic acid residues (Fig. 6). This is an interesting result, since substitutions of charged residues are not frequent mutations among homologous proteins [44] and we hypothesize that such charge differences are the main divergence among TrMSP and all other TrypMSPs studied here. Indeed, when

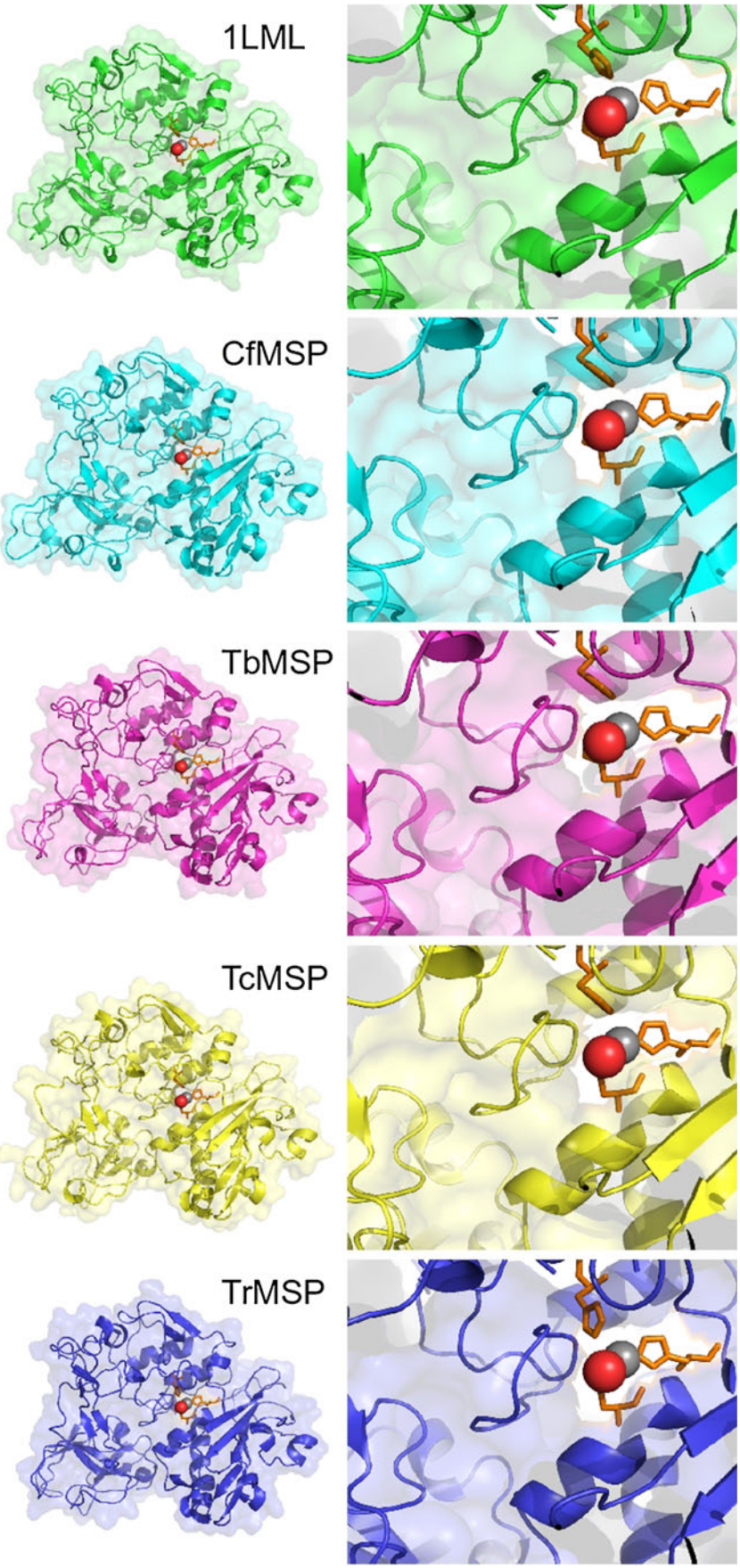

Fig. 4 Structural comparison among TrypMSPs from *Leishmania* (1LML, green), *Chritidia fasciculata* (*cyan*), *T. brucei* (*magenta*), *T. cruzi* (*yellow*) and *T. rangeli* (*blue*). Catalytic regions are enlarged for a more detailed view. Zinc ions are represented as *gray spheres* and the conserved water molecule as *red spheres*. Catalytic residues coordinating the zinc ion are represented in licorice and colored in *orange*

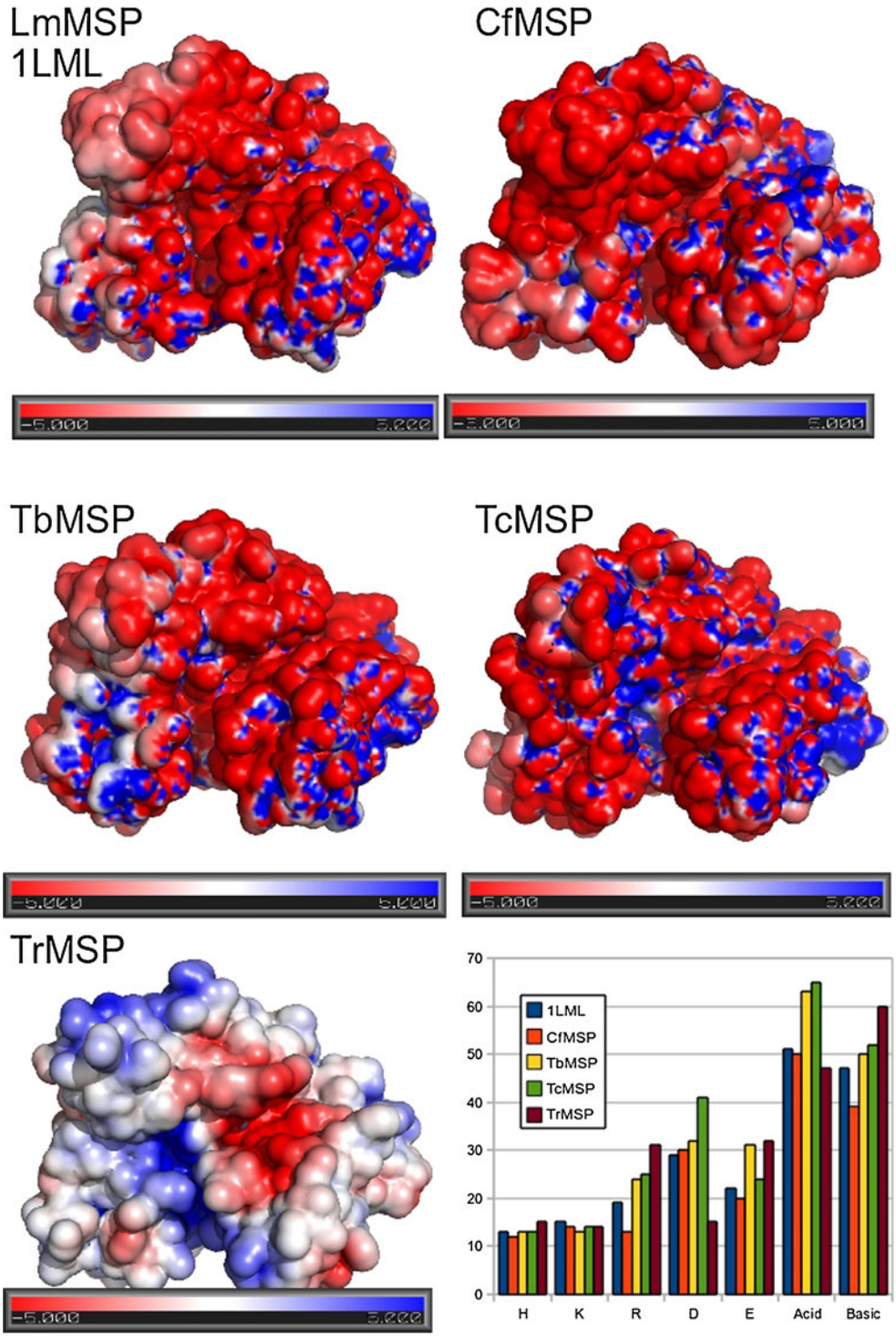

Fig. 5 Analysis of surface charges of TrypMSPs. Structures of TrypMSPs from *Leishmania* (1LML), *Chritidia fasciculata*, *T. brucei*, *T. cruzi* and *T. rangeli* showing the charge distribution in protein surfaces. Solvent-accessible residues are colored according to their charge as described in the scale below each structure (varying from more negative, in *red* to more positive, in *blue*). The histogram indicates the amount of charged residues in each protein: H, K and R are basic (positive) residues, while D and E are acid (negative) residues

consulting the BLOSUM62 matrix [45], which is the reference for assessing the probability of amino acids substitutions between homologous proteins, we observe that substitutions from negative to non-negative residues and from non-positive to positive are very unlikely in sequences that have up to 62 % identical residues (TcMSP and TrMSP have 58 % identical residues). BLOSUM62 displays the log-odds for each possible amino acid substitution, where positive scores indicate conservative substitutions and negative scores indicate non-conservative substitutions [46]. Events that substitute non-positive residues for arginine or lysine residues have an average log-odd of −2, while substitutions of non-positive residues for histidine have an average log-odd of −0.5. Considering the substitutions from negative to non-negative residues, when aspartic acid or glutamic acid residues are substituted by any non-negative

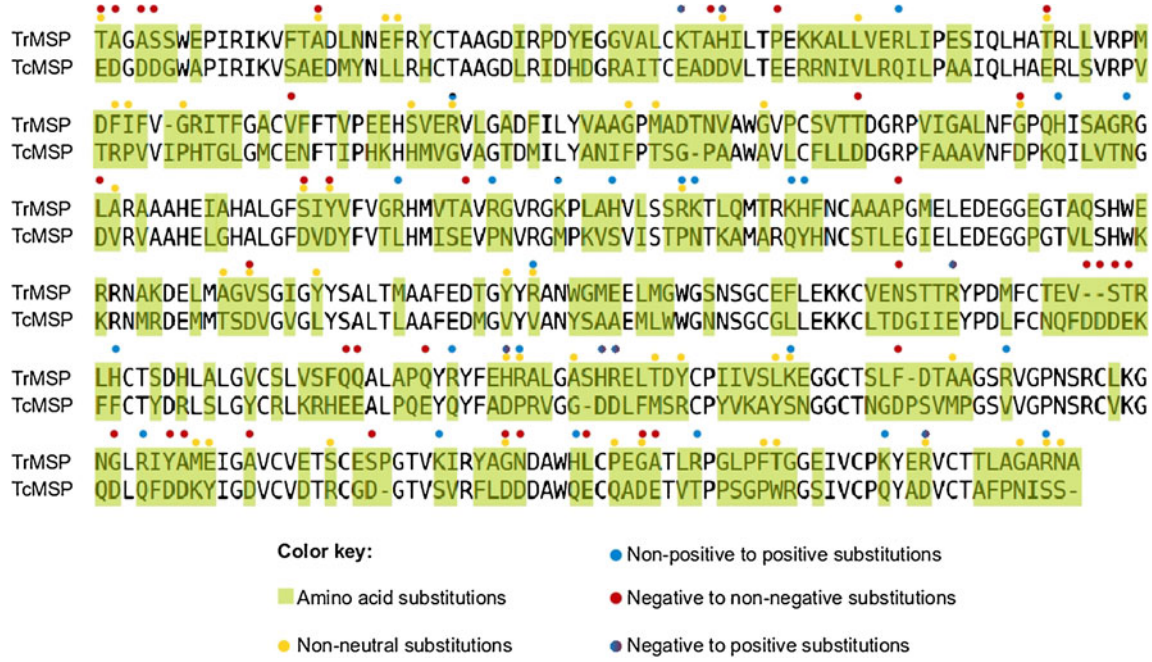

Fig. 6 Pairwise sequence alignment between TrMSP and TcMSP. All amino acid substitutions are shaded in *green* and those classified as non-neutral by SNAP are marked with a *yellow dot*. Substitutions that change a non-positive residue (in TcMSP) to a positive one (in TrMSP) are marked with a *blue dot*, while substitutions that change a negative residue (in TcMSP) to a non-negative one (in TrMSP) are marked with a *red dot*. Substitutions that change a negative residue (in TcMSP) to a positive one (in TrMSP) are marked with a double-colored dot (*red and blue*). The later are potential targets for further protein engineer studies by site-directed mutagenesis to confirm their crucial importance for parasite-host interactions. These substitutions of charged residues are responsible for the more neutral (slightly positive) charges of TrMSP when compared to its closer homologue TcMSP

amino acid, the mean log-odd is −2 and −1.5, respectively. Taken together, these data indicate that the substitutions leading to the more positive charge presented by TrMSP go against the odds. This could be interpreted as a pressure forcing TrMSP to have a positive surface, maybe as a result of co-evolution between *T. rangeli* and its hosts.

Conclusions

The complete sequence of a *T. rangeli* MSP gene as well as the amino acid sequence of the predicted protein was determined. Three-dimensional models of mature TrMSP and its close homologues from several another trypanosmatid species (*C. fasciculata*, *T. brucei*, *T. cruzi*) were successfully built by homology modeling using LmMSP (PDB code 1LML) as structural template. TrypMSPs are highly conserved among other trypanosomatid species at the sequence level and the generated three-dimensional structures reveal an additional conservation at the structural level among the modeled structures and between these and the experimentally solved LmMSP structure (1LML). Interestingly, the major difference between TrMSP and other TrypMSPs refers to the surface partial charges. TrMSP has predominantly positive surface charge, whereas LmMSP, CfMSP, TbMSP and TcMSP have negative surface charges. We have evaluated the amino acid differences among TrMSP and its closest homologue (in this study) TcMSP. From all 257 positions where substitutions occur, 45 were assigned as non-neutral, indicating that these changes may be involved in functional differences between TcMSP and TrMSP. Protein engineering studies can be further performed to confirm these predictions and characterize TcMSP sites that can account for *T. cruzi* pathogenicity to humans.

MSPs are important surface proteins that play an important role in the host-parasite interaction in trypanosomatids. Different from *T. brucei* and *T. cruzi*, *T. rangeli* is not pathogenic to humans and this may be (at least partially) due to differences in MSPs. Considering these observations, the unique characteristics of TrMSP could help to explain the lack of parasitism by *T. rangeli* and further studies may elucidate the significance of our results for protein interactions of MSPs. These results open new possibilities to the study of *T. rangeli* as a model organism for the investigation of the molecular basis of parasitism in trypanosomatids and drive attention to MSP molecules and the differences among them as a key determinant for pathogenicity.

Acknowledgments Authors thank the members of the Laboratório Multiusuário—UFTM for processing DNA sequences. This work was supportted by FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) and CAPES (Coordenação de Aperfeiçoamento e Capacitação de Pessoal de Nível Superior). PHMC was recipient of a Master fellowship from CAPES.

References

1. Mathers CD, Ezzati M, Lopez AD (2007) Measuring the burden of neglected tropical diseases: the global burden of disease framework. PLoS Negl Trop Dis 1(2):e114
2. de Moraes MH, Guarneri AA, Girardi FP, Rodrigues JB, Eger I, Tyler KM, Steindel M, Grisard EC (2008) Different serological cross-reactivity of *Trypanosoma rangeli* forms in *Trypanosoma cruzi*-infected patients sera. Parasit Vectors 1(1):20
3. Etges R, Bouvier J, Bordier C (1986) The major surface protein of *Leishmania* promastigotes is a protease. J Biol Chem 261(20):9098–9101
4. McGwire BS, O'Connell WA, Chang KP, Engman DM (2002) Extracellular release of the glycosylphosphatidylinositol (GPI)-linked *Leishmania* surface metalloprotease, gp63, is independent of GPI phospholipolysis: implications for parasite virulence. J Biol Chem 277(11):8802–8809
5. Schlagenhauf E, Etges R, Metcalf P (1998) The crystal structure of the *Leishmania major* surface proteinase leishmanolysin (gp63). Structure 6(8):1035–1046
6. Brittingham A, Morrison CJ, McMaster WR, McGwire BS, Chang KP, Mosser DM (1995) Role of the *Leishmania* surface protease gp63 in complement fixation, cell adhesion, and resistance to complement-mediated lysis. J Immunol 155(6):3102–3111
7. Hey AS, Theander TG, Hviid L, Hazrati SM, Kemp M, Kharazmi A (1994) The major surface glycoprotein (gp63) from *Leishmania major* and Leishmania donovani cleaves CD4 molecules on human T cells. J Immunol 152(9):4542–4548
8. Mosser DM, Edelson PJ (1985) The mouse macrophage receptor for C3bi (CR3) is a major mechanism in the phagocytosis of *Leishmania* promastigotes. J Immunol 135(4):2785–2789
9. Jaramillo M, Gomez MA, Larsson O, Shio MT, Topisirovic I, Contreras I, Luxenburg R, Rosenfeld A, Colina R, McMaster RW, Olivier M, Costa-Mattioli M, Sonenberg N (2011) *Leishmania* repression of host translation through mTOR cleavage is required for parasite survival and infection. Cell Host Microbe 9(4):331–341
10. Grandgenett PM, Otsu K, Wilson HR, Wilson ME, Donelson JE (2007) A function for a specific zinc metalloprotease of African trypanosomes. PLoS Pathog 3(10):1432–1445
11. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Van Aken S, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309(5733):409–415
12. Cuevas IC, Cazzulo JJ, Sanchez DO (2003) gp63 homologues in *Trypanosoma cruzi*: surface antigens with metalloprotease activity and a possible role in host cell infection. Infect Immun 71(10):5739–5749
13. Lescrinier E (2011) Structural biology in drug development. Verh K Acad Geneeskd Belg 73(1–2):65–78
14. Ferreira KA, Ruiz JC, Dias FC, Lages-Silva E, Tosi LR, Ramirez LE, Pedrosa AL (2010) Genome survey sequence analysis and identification of homologs of major surface protease (gp63) genes in *Trypanosoma rangeli*. Vector Borne Zoonotic Dis 10(9):847–853
15. Sambrook J, Russell DW (2006) The condensed protocols from Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
16. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340(4):783–795
17. Pierleoni A, Martelli PL, Casadio R (2008) PredGPI: a GPI-anchor predictor. BMC Bioinformatics 9:392
18. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF (1999) Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 112:531–552
19. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
21. Pei J, Tang M, Grishin NV (2008) PROMALS3D web server for accurate multiple protein sequence and structure alignments. Nucleic Acids Res 36 (Web Server issue):W30–W34
22. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33 (Web Server issue):W36–W38
23. Sali A, Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815
24. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK—a program to check the stereochemical quality of protein structures. J App Cryst 26:283–291
25. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35 (Web Server issue):W407–W410
26. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 98(18):10037–10041
27. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KMJ, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197
28. Li H, Robertson AD, Jensen JH (2005) Very fast empirical prediction and interpretation of protein pKa values proteins. Proteins 61:704–721
29. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res 32:665–667
30. The PyMOL Molecular Graphics System. Schrödinger, LLC., 1.5.0.4 edn
31. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35(11):3823–3835
32. Chaudhuri G, Chang KP (1988) Acid protease activity of a major surface membrane glycoprotein (gp63) from *Leishmania mexicana* promastigotes. Mol Biochem Parasitol 27(1):43–52
33. Miller RA, Reed SG, Parsons M (1990) *Leishmania* gp63 molecule implicated in cellular adhesion lacks an Arg-Gly-Asp sequence. Mol Biochem Parasitol 39(2):267–274
34. Yao C, Donelson JE, Wilson ME (2003) The major surface protease (MSP or GP63) of *Leishmania* sp. Biosynthesis, regulation of expression, and function. Mol Biochem Parasitol 132(1):1–16
35. Steinkraus HB, Greer JM, Stephenson DC, Langer PJ (1993) Sequence heterogeneity and polymorphic gene arrangements of the *Leishmania guyanensis* gp63 genes. Mol Biochem Parasitol 62(2):173–185
36. Laskowski RA (2009) PDBsum new things. Nucleic Acids Res 37: D355–D359

37. Gomis-Ruth FX (2003) Structural aspects of the metzincin clan of metalloendopeptidases. Mol Biotechnol 24(2):157–202
38. Razzazan A, Saberi MR, Jaafari MR (2008) Insights from the analysis of a predicted model of gp63 in *Leishmania donovani*. Bioinformation 3(3):114–118
39. Castrignano T, De Meo PD, Cozzetto D, Talamo IG, Tramontano A (2006) The PMDB Protein Model Database. Nucleic Acids Res 34 (Database issue):D306D–309
40. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran AN, Wortman JR, Alsmark UC, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N (2005) Comparative genomics of trypanosomatid parasitic protozoa. Science 309(5733):404–409
41. Joshi PB, Kelly BL, Kamhawi S, Sacks DL, McMaster WR (2002) Targeted gene deletion in *Leishmania major* identifies leishmanolysin (GP63) as a virulence factor. Mol Biochem Parasitol 120(1):33–40
42. Sheinerman FB, Norel R, Honig B (2000) Electrostatic aspects of protein–protein interactions. Curr Opin Struct Biol 10(2):153–159
43. Inverso JA, Medina-Acosta E, O'Connor J, Russell DG, Cross GA (1993) *Crithidia fasciculata* contains a transcribed leishmanial surface proteinase (gp63) gene homologue. Mol Biochem Parasitol 57(1):47–54
44. Barnes MR (2007) Bioinformatics for geneticists: a bioinformatics primer for the analysis of genetic data, 2nd edn. Wiley, Chichester
45. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919
46. Eddy SR (2004) Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol 22:1035–1036