Structural Modelling and Characterization of Target Specifics of *Trypanosoma cruzi*, Etiologic agent of Chagas Disease.

Présentée et Soutenue Publiquement le 16 Decembre 2016

Par
Carlyle Ribeiro Lima

Dirigée Par
Pr. P. Derreumaux et Dr. P. Tufféry

Jury

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philippe Derreumaux</td>
<td>Pr. Univ Paris Diderot</td>
</tr>
<tr>
<td>Pierre Tufféry</td>
<td>DR. INSERM</td>
</tr>
<tr>
<td>Jerome Golebiowski</td>
<td>Pr. Univ. de Nice</td>
</tr>
<tr>
<td>Jean François Zagury</td>
<td>Pr. CNAM Paris</td>
</tr>
<tr>
<td>Patrick Fuchs</td>
<td>MCF Univ Paris Diderot</td>
</tr>
<tr>
<td>Joel Pothier</td>
<td>MCF Univ. P. et M. Curie</td>
</tr>
</tbody>
</table>

Directeur de Thèse

Rapporteur

Rapporteur

Examineur

Examineur
Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration.

Name: Carlyle RIBEIRO LIMA

Signature: .................................. Date: .........................

Carlyle Ribeiro Lima
16 December 2016
Acknowledgements

First, I would like to express my sincere gratitude to my supervisors Prof. Philippe DERREUMAUX and Dr. Pierre TUFFERY for the continuous support of my Ph.D studies and related research, for their patience, motivation, constructive feedback, particularly during those times where I felt defeated by my research. For sure they have contributed to my training significantly, not only a scientist but also as a person. I could not have imagined having a better supervisor and mentor for my Ph.D studies.

I would like to express my gratitude also to Nicolas Carrels, Carolina Guimaraes and Ernesto Caffarena, for contribution of this journey Ph.D, Their holdings were decisive for my change of life in these three years.

I would like to express my gratituto to my family, the immense support, caring, motivation and confidence. But especially my: Grandfather, grandmother, mother and aunt, for without the support of these 4 people, would not be half the man I became today. Unfortunately my esteemed grandfather is not be able to view the end of this trajectory, but I know he would be happy to see the end of this story. AND IS HE TO WHOM I DEDICATE THIS THESIS.

I would also like to thank the friends I made here: Tristan, Maria, Mara, Marina, Thuy, Dario and Stephanie. Thank you for the conversation, happy days and learning. I would also like to thank the friends I made here: Tristan, Maria, Mara, Marina, Thuy, Dario and Stephanie. Thank you for the conversation, happy days and learning. To friends who are in Brazil, is also my thanks because every holiday is a renewal I able to rely with you and know that they will be all together when I return.

It was a grateful opportunity to work at the Laboratory of Theoretical Biochemistry and Laboratory Molecules Therapeutic in Silic where I learned much during my time here, and met many great researchers and students. I would like to thank all people in the lab. Many thanks go to Victoria, secretary of IBPC, who always supportid me in the administrations and Geoffrey Letessier who help about information technologies. My project could not be finish without their help.
Abstract

According to the World Health Organization, 21 Latin American countries are endemic for Chagas disease, affecting 10 million people. Chagas’ disease, caused by the protozoan Trypanosoma cruzi (T. cruzi), is a parasitic illness endemic mostly in Latin America and particularly in Brazil. Despite many experimental studies, there is no efficient treatment against Chagas disease, and the search for new therapeutic targets specific to T. cruzi is critical for drug development. In my thesis, I have revisited 41 protein sequences proposed by the analogous enzyme pipeline, and found that it is possible to provide structures for 33 T. cruzi sequences with clear homologs or analogs in H. sapiens and likely associated with trypanothione reductase, cysteine synthase and ATPase functions, and structures for sequences specific to T. cruzi and absent in H. sapiens associated with 2,4-dienoyl-CoA reductase, and leishmanolysin activities. The implications of our structures refined by atomistic molecular dynamics (monomer or dimer states) in their in vitro environments (aqueous solution or membrane bilayers) are discussed for drug development and suggest that all protein targets, except cysteine synthase, merit further investigation.
Résumé

Selon l'organisation mondiale de la santé, 21 pays d'Amérique Latine et notamment le Brésil sont touchés par la maladie de Chagas qui est provoquée par le protozoaire Trypanosoma cruzi (T. cruzi). En dépit de très nombreuses études expérimentales, il n'existe aucun traitement efficace contre la maladie de Chagas, et l'identification de nouvelles cibles thérapeutiques est fondamentale pour développer de nouvelles drogues. Dans cette thèse, j'ai revisité 41 protéines chez T. cruzi récemment proposées par Nicolas Carels et al. A partir d'une nouvelle procédure bioinformatique, j'ai trouvé qu'il est possible de prédire les structures 3D de séquences T. cruzi ayant des homologues ou des analogues chez l'homme et très probablement associées aux fonctions trypanothione réductase, cystéine synthase et ATPase, ainsi que des séquences spécifiques à T. cruzi et absentes chez Homo Sapiens associées à des activités 2,4-dienoyl-CoA réductase et leishmanolysine. Les implications de nos modèles raffinées par des dynamiques moléculaires atomistiques (monomère ou dimère) dans leur environnement in vitro (solution aqueuse ou bicouche membranaire) sont discutées à des fins thérapeutiques, et suggèrent que toutes les cibles protéiques, à l'exception de la cystéine synthase, méritent d'autres investigations.
## Contents

**Chapitre I** – la Maladie de Chagas. .......................................................... 7
**Les Maladies Tropicales Negligées (MTN).** .................................................. 8
**La Maladie de Chagas.** ............................................................................. 9
  - Pathogenèse et manifestations cliniques. ............................................. 10
  - Une large diffusion au niveau mondial. ........................................... 13
**Trypanosoma cruzi (Tc).** ......................................................................... 18
  - Morphologie de Trypanosoma cruzi. ................................................. 18
  - Cycle de l'infection. ........................................................................... 20
  - La ligne de défense immunitaire contre T. cruzi. .............................. 25
  - La Rupture des Cellules. ................................................................. 28
  - Les cellules T-CD+ 4 et 8 ............................................................... 30
**Chapitre II** .............................................................................................. 40
**Traitements existants contre la maladie de chagas.** .................................. 41
**Solutions thérapeutiques en cours d'exploration pour le traitement de la maladie de Chagas.** ............................................................... 43
  - Peptides antimicrobiens - Gramicidine A. ......................................... 43
  - Peptides de venin - Méttitine ........................................................... 44
  - Dérivés de phthalazine ...................................................................... 47
  - Terpenes - Acide bêtulinique ............................................................. 50
**La recherche de nouvelles cibles thérapeutiques contre la maladie de Chagas.** .................................................................................... 52
  - Thiol Transferase (TC52) ................................................................ 52
  - Cruzipaïne 1 et 2 ............................................................................... 53
  - Trans-sialidase de Trypanosoma cruzi (TS) ....................................... 54
  - Glycéraldéhyde 3-phosphate déshydrogénase (GAPDH) ................. 55
  - Sterol 14α-demétilase (CYP51) ......................................................... 55
**Conclusion** .............................................................................................. 56
**CHAPITRE III** ......................................................................................... 63
**Aims of the Thesis** .................................................................................. 64
**CHAPITRE IV** .......................................................................................... 68
**Materials and Methods** ........................................................................ 69
**CHAPITRE V** .......................................................................................... 79
**Article** ..................................................................................................... 80
**CHAPITRE VI** .......................................................................................... 118
**Conclusions** ........................................................................................... 119
**Annexe** ................................................................................................... 123
la Maladie de Chagas.
la Maladie de Chagas.

**Les Maladies Tropicales Negligées (MTN).**

Les maladies tropicales négligées (MTN) ont une incidence sur un peu plus de 1,4 milliard de personnes dans le monde, dont plus de 800 millions d'enfants à faible revenu et faible qualité de vie. En plus du large spectre d'infection chez les personnes à faible revenu, on estime qu'elles peuvent causer plus de 500.000 morts par an, et engendrer malnutrition et déficiences cognitives [1,2]. Le terme des maladies tropicales négligées (MTN) a été proposé d'abord dans les années 1970 par Ken Warren [3], qui a porté son attention sur une grande partie de la population humaine possédant un faible pouvoir économique et souffrant de maladies chroniques, invalidantes et non fatales, avec une faible attractivité financière pour les gouvernements et les industries pharmaceutiques. La lutte contre certaines MTNs (Special Programme for Research and training in Tropical Diseases, Medicines for Chagas Diseases, Drugs for Neglected Diseases initiatives), a progressé de façon régulière depuis l'injection de ressources financières et technologiques par des partenaires internationaux après la Déclaration de Londres en 2012 [4]. Cette déclaration a initié un plan avec des objectifs détaillés pour chacune de ces maladies. Parmi les grands projets, la promotion de conditions d'hygiène minimales et de l'usage de l'eau propre ainsi qu'une amélioration de la qualité de vie des populations les plus pauvres et les plus reculées dans les grands centres urbains permettraient de mieux contenir la prolifération des maladies [4].

Les succès des programmes sont mesurés par un grand nombre d'indicateurs comme le nombre de zones à risque, le nombre de cas enregistrés, la distribution des médicaments, la fréquence du traitement, la durée du traitement, l'amélioration de l'indice de développement humain, etc. Mais il faut noter que même si les indicateurs évoluent favorablement, des différences notables entre les objectifs initiaux et les objectifs atteints persistent [5,6]. Ainsi, la reconnaissance croissante de la nécessité de redoubler d'efforts dans la lutte contre les maladies non transmissibles, et la mise en place de politiques visant à améliorer la qualité de vie de la population à risque ont résulté en une collaboration entre les différents secteurs de la société (public-privé). Il a notamment été pointé qu'un plus grand accès des médicaments aux patients, non seulement durant les phase symptomatique et chronique, les campagnes ciblées mais aussi et à tout moment de l'infection, donnerait un meilleur contrôle sur les MTNs [7].
La Maladie de Chagas.

La maladie de Chagas ou trypanosomiase américaine, est une maladie endémique qui affecte les pays d'Amérique latine, d'Amérique centrale et d'Amérique du nord. C'est une maladie infectieuse décrite en détail plus loin - transmise par les espèces vectrices de triatomes de la famille Reduviidae, populairement connue comme “barbeiro” (Figure 1A) qui résidaient sur les murs des maisons "pau-a-pique ou en torchis" (Figure 1C) [8,9]. La maladie de Chagas a été découverte en 1907 par le chercheur Carlos Justiniano Ribeiro Chagas en raison d'une suspicion de malaria dans la ville de Lassance dans le Minas Gerais (Brésil). Il a identifié une nouvelle espèce de parasite (figure 1B), qui jusque-là était nommé *Schyzotrypanum cruzi*, puis a été rebaptisé plus tard *Trypanosoma cruzi* [10,11]. Après l'identification de l'agent infectieux, les symptômes et les anomalies ont été associées (croissance des muscles de certains organes spécifiques, le cœur, l'œsophage et le côlon) à la présence du parasite.

*Figure 1: A) Vecteur de transmission: Triatoma infestans. B) Forme de l'agent étiologique: Trypanosoma Cruzi. C) La maison insalubre type abritant la reproduction du vecteur et la transmission.*

la Maladie de Chagas.

La maladie de Chagas est une maladie infectieuse considérée comme une MTN par l'OMS car elle est étroitement liée à de mauvaises conditions de développement socio-économique, et est associée à un taux élevé de la mortalité dans les pays endémiques [12,13]. Initialement, la maladie était confinée aux continents américains, lui attribuant ainsi le nom “Tripanossomiase americana”, et il existait une gamme de plus de 140 types de insectes vecteurs de maladies [14].

Pathogenèse et manifestations cliniques.

Comme l’illustre la Figure 2, l’entrée du parasite dans l’hôte vertébré peut engendrer des lésions comme le Chagome (Figure 2 : à droite et durée de 2 semaines ou plus) qui correspond à une infection au niveau de la piqure. Quand la piqure est localisée dans la région oculaire les lésions sont appelées Roman (Figure 2 : à gauche et type d’œdème unilatéral bipalpebral, adenite préauriculaire). Dans environ 90% des cas le parasite est éliminé par les défenses naturelles de l’hôte, sans utilisation de médicaments [15,16]. Lorsqu’elle se déclare, la maladie de Chagas comporte ensuite deux phases: Aigüe et Chronique. La phase aigüe est associée à un nombre important de parasites dans le sang, ce qui favorise la diffusion dans les organes et l’infection tissulaire. Cette phase est courte chez H.sapiens, de l’ordre de 30 à 60 jours, avec des symptômes divers : fièvre de 1 à 2 semaines, hépatosplénomégalie, adénopathies, myocardite, électrocardiogramme modifié, oedèmes subcutanés ou généralisés. Les symptomes dépendent de la parasitémie de l’individu infecté qui peut varier en fonction du lieu de la contamination, et la nature du triatomine contaminant [17,18]. Dans la phase aigüe, le risque d’affection du myocarde est grand, souvent sous la forme de myocardite aiguë diffuse, conduisant à la mort.

Figure 2: Résultat d'une infection par vecteur de la maladie de Chagas: Roman (à gauche) et chagome (à droite). (Images en ligne : http://anotacoesdalua.blogspot.fr/2010_12_01_archive.html)

Après un pic de parasites dans le corps, les symptômes de la maladie ont tendance à régresser pendant une phase de latence qui se caractérise par une réduction du nombre
la Maladie de Chagas.

de parasites dans le corps. Cette phase évoluera dans environ 50% - 70% des cas en phase chronique de la maladie de Chagas. Parmi ces cas, environ 20 à 30% déclarent une phase chronique avec des symptômes individuels bien identifiés comme des maladies cardiaques, mégacôlon et megaesophagus (Figure 3) [19-21]. L’affection du tractus gastro-intestinal (TGI) est une manifestation clinique de la maladie de Chagas, qui est présente dans 10% à 15% des cas chroniques, où il est possible de noter une augmentation du volume des organes qui composent ce système, ainsi que des modifications de la motilité, la douleur, la régurgitation, l’hypertrophie des muscles lisses et l’atrophie de la muqueuse, dues à la destruction de neurones du plexus muqueux par le parasite et une défaillance de la réponse immunitaire du corps. La figure 3 montre les différentes manifestations provoquées par la maladie de Chagas dans le système TGI. Cependant, la relation de la maladie de Chagas avec des manifestations d’ordre digestif (Colon et l’oesophage), reste débattue [22,23].

la Maladie de Chagas.

Les formes les plus mortelles de la maladie de Chagas, sont associées à des cardiopathies qui conduisent à la mort subite (66% des cas), suivie de près par l'insuffisance cardiaque et de thromboembolie [24]. Cette forme de la phase chronique de la maladie de Chagas est associée à une inflammation évolutive du myocarde, qui est progressivement entouré d'un réseau de fibrine qui tend à détruire le muscle cardiaque, provoquant des symptômes tels que l'arythmie, l'extension des parois du myocarde [25,26]. Le cœur peut subir des modifications de taille et de poids, et ces changements peuvent être expliqués par deux hypothèses: i) des lésions dans le tissu cardiaque résultant des réponses immunitaire de l'organisme à la présence du parasite. ii) des dommages des cellules cardiaques et notamment des cellules du tissu du myocarde [22,25,26]. La figure 4 illustre les principaux problèmes cardiaques résultant d'une infection par T. cruzi.

Figure 4: Common findings in chronic Chagas heart disease. (A) forme segmentaire cardiaque. (B) forme cardiaque dilatée globale.

* Tous les résultats présentés pour les patients avec la forme segmentaire l'cardiaque peuvent également survenir chez les patients atteints de la forme dilatée mondiale. LV = ventricule gauche. RV = ventricule droit. LA = oreillette gauche. RA = oreillette droite. (D'après Rassi et al, 2012).

Dans le cas d'une co-infection par le VIH, le comportement du parasite de la maladie du Chagas est celui d'un microorganisme opportuniste, ce qui complique le diagnostique. Même si les premiers rapports de co-infection datent des années 90, la
la Maladie de Chagas.
fréquence de cette association évolue, et provoque une augmentation du nombre de décès et amplifie globalement la recherche d’un traitement de la maladie de Chagas à l’échelle mondiale [27-29].

**Une large diffusion au niveau mondial.**

En raison de l’augmentation de l’immigration dans de nombreux pays à travers le monde, la propagation de la maladie est en passe de devenir un problème de santé dans les régions de l’Amérique du Nord, en Europe et en Asie [30]. La répartition et l’estimation de l’infection dans les différentes régions du monde sont présentées (Figure 5). Des études récentes montrent qu’il y a environ 8 millions de personnes atteintes par la maladie de Chagas dans le monde, avec des incidences les plus fortes dans la zone allant du Mexique à l’Argentine, en raison de la parasitémie (quantité de parasite présente dans le sang humain) endémique provoquant une exposition de l’ordre de 75 à 90 millions de personnes [29]. Malgré la variation des estimations réalisées par différents groupes de recherche, ces nombres sont considérables, en particulier dans les pays où les conditions sanitaires sont précaires, et requièrent une plus grande attention des autorités politiques.

La migration (régulière ou clandestine) accrue au niveau mondial a augmenté la possibilité d’infection dans les pays jusque-là peu endémiques, impliquant des régions de l’Amérique du nord, de l’Europe, de l’Asie et même de l’Océanie [31].

![Figure 5: Distribution internationale estimée des cas de la maladie de Chagas. (D’après WHO, 2015)](image)

En Europe, le problème est aggravé dans les pays à forte migration. Des pays comme la France, l’Espagne et le Royaume-Uni (Figure 6) ont mis en place un programme
la Maladie de Chagas.
de dépistage des poches de sang qui sont dans leurs inventaires, ce qui peut faciliter la détection précoce de la maladie, éviter la diffusion, et permettre le traitement des donneurs de sang contaminés [32-34].

Figure 6: Distribution estimée des cas de la maladie de Chagas en Europe (D'après Kichen et al 2015).

Au cours des dernières années, il y a eu une augmentation des flux migratoires internationaux en raison des crises économiques et des guerres dans les pays plus pauvres (pays latins, africains et du Moyen-Orient), ce qui facilite la propagation de la maladie de Chagas dans les pays non endémiques [34,35].

En Europe, les migrants ont souvent accès à des ressources financières abondantes par rapport à leurs pays d'origine ce qui a augmenté l'attractivité des pays Européens. Cette attractivité a été couplée à un flux migratoire et des activités légales ou illégales [34,35], qui rendent le décompte exact du nombre de personnes qui entrent dans l'Europe, et en conséquence le dépistage de la maladie difficiles. De plus, en raison de la crise économique, un grand nombre d'immigrants est soit retourné dans leur pays d'origine ou soit a migré vers d'autres pays Européens, ce qui est facile dans les pays signataires des accords de Schengen. Ces mouvements incontrôlés sans dépistage des personnes...
la Maladie de Chagas.

infectées peuvent conduire à une diffusion massive de la maladie de Chagas. La mesure de cette migration en Europe n'étant pas obligatoire, les données pour les 1-2 dernières années sont quasi-inexistantes [32-35].

Aux États-Unis, le problème est aggravé par la grande quantité de Triatominae. Comme le montré la Figure 7, environ 10 espèces de Triatominae sont présents dans 2/3 des États-Unis, avec une prédominance dans la partie la plus méridionale [36].

![Figure 7: Répartition des Triatominae – vecteur des parasites de la maladie de Chagas aux USA (D'après Klotz et al, 2015)](image)

À l'exception de l'espèce Triatominae rubrofasciata, le reste des Triatominae est présent au Mexique. Les variations peuvent être attribuées à la mutation des espèces au fil du temps, au degré de contrôle et de surveillance de certains états, et à certaines erreurs d'identification des espèces dans les différents états [37,38]. A la différence de certains insectes de l'Amérique centrale et du sud, certaines de ces espèces ne sont pas des de bons vecteurs pour la propagation de la maladie de Chagas, car ces insectes se nourrissent du sang des poulets, qui ne constituent pas un réservoir de T. Cruzi [36,39]. Leurs températures corporelles élevées et leurs mécanismes de défenses naturelles ne sont pas compatibles avec la présence du parasite vivant. Cette situation ne génère pas une grande inquiétude au sujet d'une éventuelle épidémie de la maladie de Chagas aux États-Unis. En plus de la diffusion restreinte de T. Cruzi, la qualité de vie de ses habitants...
la Maladie de Chagas.
est généralement élevée et ne favorise pas une propagation de l'épidémie [36,39,40].

En Amérique latine, l'infection se produit généralement par Triatoma infestans qui se nourrit de sang humain, avec des caractéristiques nocturnes, et une large adaptation aux environnements humains, comme les murs fissurés ou les plafonds et les planchers des logements précaires. On estime à plus de 1 million les cas d'infection depuis 2010 dans des pays tels que le Brésil et l'Argentine. Cependant, des taux d'infection plus élevés sont observés dans la plupart des pays des régions amazoniennes ( Pérou, Brésil, Bolivie, Equateur, Colombie, Venezuela, Guyana et Guyane française) [41].

Au Brésil aujourd'hui, il y a une prédominance des cas chroniques de la maladie de Chagas, causée par des infections acquises il y a plus de 15 ans, avec une prévalence de 3.000.000 d'individus infectés dans les zones d'endémie étudiées pendant plus de 30 ans. Au cours des 10 dernières années, le nombre de cas de la maladie de Chagas en phase aiguë a été croissant dans l'Amazonie brésilienne (ou légale, comprenant 9 états sur les 27 états du pays), et semble plus stable dans les zones endémiques des 18 autres états [30]. Pour la région de l'Amazonie juridique, un point qui a attiré l'attention des chercheurs sur la maladie de Chagas est l'augmentation du nombre de vecteurs de la maladie de Chagas en liaison avec une transmission supplémentaire impliquant l'ingestion d'aliments contaminés (jus de canne, palmier pinot, parmi d'autres). Une analyse portant sur un millier de cas de ces nouveaux modes de transmission a montré que 70% ont eu lieu par la transmission par voie orale, 7% par la transmission du vecteur et les derniers 23% restent indéterminés [30].

La diminution du nombre d'études systématiques sur une base démographique fait obstacle à une meilleure compréhension de l'importance de la maladie de Chagas, et son évolution au cours du temps. Dans les années 50-60, on pensait la maladie limitée à la zone rurale, avec transmission élevée par Triatoma infestans. Durant la décennie suivante qui correspond à l'industrialisation au Brésil, de nombreux habitants des régions rurales ont migré vers les grandes villes, mais dans conditions encore plus précaires que les zones rurales, rendant la cartographie et le contrôle de la maladie de Chagas quasi-inexistants dans la zone urbaine [12].

En conclusion, même après un peu plus de 106 ans, et des efforts importants au niveau politique, technique, et scientifique, il reste de grands efforts à accomplir. La dimension continentale du Brésil (8.516.000 Km²), et ses fortes inégalités sociales ne
la Maladie de Chagas.

permettent pas un contrôle efficace de la maladie de Chagas. Les zones les plus éloignées des grandes villes (axe Nord-Nordest) sont des zones critiques pour la transmission de cette maladie, car leur population reste vulnérable [29].
la Maladie de Chagas.

**Trypanosoma cruzi (Tc).**

**Morphologie de Trypanosoma cruzi.**

Le Trypanosoma cruzi représenté en Figure 8 appartient la famille Trypanosomatidae, de l'ordre de Kinetoplastida (Euglenozoa, Kinetoplastea). Les Trypanosomatidae sont des protozoaires à un seul flagelle, trouvés principalement dans les insectes de la famille des Reduviidae. Leur cellule contient une mitochondrie qui s'étend très largement. Près du corps basal, il y a le kinetoplaste, un arrangement composé de fibres d'ADN, aussi appelé kDNA. T.cruzi, a un génome diploïde d'environ 22.570 gènes – le génome complet de T. Cruzi n'est pas encore séquencé - qui codent pour des protéines, dont beaucoup ne sont pas identifiées (indéterminées) ou correspondent à des protéines hypothétiques. Les protéines identifiées les plus représentatives avec un grand nombre de séquences chez T. cruzi sont: des peptidases, des transporteurs ABC, des protéines de choc thermique, des ADN et ARN hélicases et des glycosyltransférases [42].

La surface cellulaire des trypanosomatides se compose essentiellement de deux parties: la membrane plasmique et une couche comportant de microtubules. La membrane plasmique de T. cruzi a une taille maximale de 10 nm et couvre toute la surface du corps cellulaire du parasite. Sous la membrane, on peut voir une mince couche de microtubules de 8 nm qui se prolongent dans le corps cellulaire. L'association membrane plasmique - microtubule est si forte dans les trypanosomes que ces deux composantes restent liées même après la lyse du protozoaire. En ce qui concerne la composition chimique, la membrane de la forme épimastigote est formée par une double couche lipidique, composée de 60% d'hydrate de carbone neutre, 12,6% de lipides, 9,5% de protéines, le reste étant constitué de 17,9% de lipophosphoglycannes (LPFG) [43].

La région d'où émerge le fléau (sac flagellaire), est formé par l'invagination de la membrane plasmique qui établit une continuité directe avec la membrane du flagelle. Comme la membrane du corps cellulaire et le fléau établissent un contact au point d'émergence du flagelle, le fléau peut être considéré comme un compartiment extracellulaire spécial qui est la seule région qui n'a pas la couche de microtubules.
La proportion de la longueur totale du flagelle varie en fonction du stade de développement. Chez T. cruzi, le filé a une longueur allant de 1 μm pour la forme amastigote à environ 20 μm pour la forme épimastigote [44] (voir la section suivante). Le flagelle est fixé au corps cellulaire.

T. cruzi comporte un noyau entouré d'une membrane poreuse, typique des eucaryotes, avec une chromatine condensée observée dans les cellules épimastigotes mais pas dans amastigoras (forme réplicative due à la fission binaire) et trypomastigotes.
la Maladie de Chagas.

Le nombre de chromosomes varie selon les trypanosomatides, il est de 41, avec des tailles allant de 78 kb à 2.4MB pour T. Cruzi [45,46].

Cycle de l'infection.

*Trypanosoma cruzi* possède plusieurs caractéristiques morphologiques et métaboliques qui permettent l'adaptation à des environnements différents. Son cycle de vie comprend quatre phases distinctes dans la littérature: épimastigotes, trypomastigotes métacycliques (tube digestif de l'insecte vecteur), amastigote (intracellulaire) et trypomastigotes (exclusive hôtes vertébrés) [47]. Il existe des formes intermédiaires, cependant moins importantes dans l'infection, tels que: *Sleder et broad* (population de tripomastigotes situé dans le sang du mammifère) et spheromastigotes (stade de différenciation des trypomastigotes sanguins et épimastigote, situé au milieu de l'intestin de l'insecte). Ces étapes peuvent être identifiées par morphométrie, la position kinétoplaste sur le nucleo et émergentes région du fléau [48].

Dans l'hôte invertébré triatomine, le cycle de vie *T. cruzi* commence après l'ingestion de sang infecté. En arrivant à l'estomac du vecteur, les trypomastigotes sanguins initient une différenciation (fission binaire) en epimastigotes. Ces formes migrent vers l'intestin où elles se multiplient. Dans la partie médiane et postérieure de l'intestin, la différenciation des cellules conduit du stade epimastigotes au stade métacycliques (formes infectieuses) qui sont libérés dans les fèces et l'urine de triatomines [44]. L'infection de l'hôte vertébré se produit par insertion des fèces du parasite dans la circulation sanguine de l'hôte, provenant de la dégradation de la peau après la piqûre de l'insecte dans les muqueuses des yeux, du nez et la bouche [49,50].

La pénétration du parasite (trypomastigotes métacycliques) dans la circulation sanguine, est suivie d'une diffusion dans certains tissus et cellules, qui sont suivis d'une différenciation. Par fission binaire, les amastigotes se reproduisent dans les cellules, et deviennent des trypomastigotes qui, en brisant les cellules infectées vont pouvoir envahir d'autres cellules [51]. Cette infection peut être divisée en trois phases: l'adhésion, la signalisation et l'invasion des cellules. L'adhésion ne conduit pas nécessairement à une invasion. L'étape d'adhérence comprend des molécules de reconnaissance présentes sur la surface des cellules du parasite et de l'hôte. Pour que l'internalisation cellulaire survienne, l'activation de voies de signalisation conduisant à l'augmentation de la concentration en calcium dans les deux organismes, de parasites et l'hôte est requise.
la Maladie de Chagas. [52,53]. Dans le cas contraire, le parasite s'échappe dans des environnements extracellulaires. La Figure 9 montre les cycles de traitement et d'infection par T. cruzi dans l'hôte vertébré et invertébré.

Il existe deux stratégies pour l'invagination cellulaire par T. cruzi. Dans la première, qui se produit dans 20% à 30% des cas, le parasite utilise la dépendance lysosomale (cellules de l'hôte), qui induit la voie de signalisation d'inositol triphosphate (IP3) qui produit une augmentation de la concentration en Ca²⁺ suivi par le recrutement de calcium par le parasite et la fusion de lysosomes de la cellule hôte par le parasite. Une autre stratégie, survenant dans 70% à 80% des cas, consisterait à l'invagination de la membrane plasmique, suivie par une fusion avec les lysosomes intracellulaires [54-57]. Ces stratégies montrent que quelle que soit la méthode adoptée par le parasite, la fusion lysosomale est essentielle à l'invasion. De plus, l'acidification de la vacuole lysosomale contribue à la différenciation en amastigotes trypomastigotes à l'intérieur de la cellule [54].

Les Trypomastigotes métacycliques peuvent se développer dans la plupart des cellules à l'exception de certains types comme les neutrophiles et les basophiles. L'entrée du parasite dans la cellule hôte est un processus complexe (Figure 10). Les trypomastigotes métacycliques entrent dans la cellule d'une manière polarisée, de
La Maladie de Chagas.

Préférence le long d’une partie de la membrane basolatérale à forte concentration en fibronectine et récepteur de la cellule hôte. Les Trypomastigotes métacycliques ont un tropisme préférentiel pour les phagocytes mononucléaires, les cellules musculaires, les cellules adipeuses et les neurones. Au bout de quelques heures dans la vacuole parasitophore, les trypomastigotes métacycliques s’échappent dans le cytoplasme où ils se différencient en amastigotes, une forme qui se réplique dans la cellule. Après division binaire intense, les amastigotes se différencient en trypomastigotes sanguins qui perturbent les cellules hôtes avec un mouvement intense des flagelles [48,58].

La Figure 10: Schéma de connexion entre trypomastigotes métacycliques et membrane cellulaire de l’hôte vertébré. (D’après Souza el al, 2010).

Le rôle des dérivés Nitrooxidatifs dans la réponse de l’hôte à T. cruzi.

T. cruzi peut être internalisé par les cellules phagocytaires, et notamment par les premiers macrophages résidant sur le site de l’infection. T. cruzi résiste à un environnement hostile, souvent oxydant à l’intérieur du phagolysosome, grâce à un réseau complexe d’enzymes antioxydantes telles que les peroxydases et les superoxyde dismutases (SOD) qui ont la fonction de protection à l’oxygène et au groupement azote réactif libéré par les macrophages [59-62]. La Figure 11 montre les reactions impliquées dans l’internalisation de T. cruzi par phagocytose.
la Maladie de Chagas.

Figure 11: Réactions favorables à l'intériorisation du Parasite dans les cellules hôtes. (A) Après la phagocytes du parasite, macrophage la NADPH oxydase associée à la membrane est activée, produisant le radical superoxyde (O$^-$2) (O2 • -) qui peut être converti en H2O2 à l'intérieur de la lumière du phagolysosome. Les macrophages stimulés par des cytokines pro-inflammatoires (IFNγ et TNF) induisent l'expression de l'oxyde nitrique synthase inductible (iNOS), générant de l'oxyde nitrique (NO • -) dans le cytoplasme de l'oxydation de la L-arginine. Le NO- diffuse dans le phagolysosome et réagit avec O2 - pour former du peroxynitrite (ONOO-), un oxydant puissant. Les radicaux libres secondaires tels que le carbonate (CO3 • -), le dioxyde d'azote (NO2 •) et hydroxyle (OH •) des radicaux, sont fabriqués à partir ONOO-. Ces espèces réactives de l'oxygène (ROS, indiquées en rouge) peuvent causer divers dommages et la mort du parasite cellulaire dans le phagolysosome. Pour survivre dans cet environnement hautement oxydant, le parasite possède un système complexe d'enzymes antioxydantes, les peroxydases (TcAPX, TcCPX et TcMPX) et superoxyde dismutase (SOD), qui agissent dans la détoxication des SOD, et sont distribuées dans différents compartiments cellulaires comme les glycosomes, la mitochondrie, le cytosol et le réticulum endoplasmique (RE). Les enzymes du glycosome, des mitochondries, du cytosol et du RE sont indiquées en orange, bleu, vert, et violet, respectivement. (B) Pour établir une infection productive, les trypomastigotes doivent migrer du phagolysosome vers le cytosol, où il se différencient en amastigotes. Dans le cytosol des macrophages, ROS, au lieu d'être préjudiciable au parasite, peut favoriser la croissance intracellulaire de T. Cruzi par le mécanisme: qui peut impliquer de faciliter l'accès de l'amastigote, au fer. Dans le cytosol, le fer peut être conservé en l'fer ferrique (Fe3+) ou peut être exporté à partir de la cellule que la ferreux (Fe2+) par la ferroportine, l'exportateur de fer particulaire des macrophages. L'expression de la ferroportine et la ferritine est régulée positivement par les antioxydants, Le mécanisme d'absorption du fer par des amastigotes est inconnu, mais le parasite peut dépendre du pool intracellulaire du fer labile pour la croissance. (D'après Cardoso et al. 2016).

Lors d'une infection par phagocytes, la présence du parasite déclenche l'activation
la Maladie de Chagas.

d'une série d'actions d'oxydation du NADPH associée à la présence de nombreux radicaux superoxydes (O₂⁻), qui peuvent être convertis en H₂O₂ par les SODs. Au cours de l'infection du parasite et la progression de la maladie de Chagas, les espèces réactives de l'oxygène (ROS) peuvent être générées dans les réactions d'immunisation médiées par suite de dommages secondaires cytotoxiques pour les mitochondries et la destruction des tissus causés par le parasite où plus tard ces ROS peuvent oxyder l'ADN, les protéines et les lipides peuvent éventuellement tuer le parasite. Un autre point de stimulation des macrophages résultant de l'infection par Trypanosoma cruzi est la production de cytokines pro-inflammatoires (IFN-γ et TNF), déclenchée suite à la grande production d'oxyde nitrique (NO) par l'intermédiaire de l'activité enzymatique de la synthase inductible de l'oxyde nitrique, qui oxydent L-arginine et transfère des électrons de NADPH [59,63-65].

La présence d'oxyde nitrique affecte le parasite, provoquant la modification chimique des protéines contenant de la cysteine, en inhibant l'activité catalytique de la cruzipâïne et la liaison avec métalloprotéines [63,66,67].

L'oxyde nitrique généré est susceptible de réagir avec l'oxygène (O₂) pour produire l'ion peroxyinitrite (ONOO⁻), un oxdant puissant et une molécule cytotoxique qui est très efficace contre T. cruzi. Cet ion peut endommager les cellules en raison de leur réaction de peroxydation lipidique, qui endommage l'intégrité des fonctions membranaires et connexes, et résulte en une altération importante de la fonction mitochondriale et induit la mort cellulaire par apoptose ou nécrose. Ces contraintes oxydatives, peuvent être nocives pour l'hôte, car peuvent présenter des risques d'endommagement des tissus, notamment par le stress nitrique, ce qui a été prouvé dans le cas de la myocardite induite lors de la progression de la maladie de Chagas [63,64].

Le système antioxydatif du parasite est constitué de diverses enzymes et des molécules enzymatiques réparties dans différents compartiments qui composent les cellules, telles que le cytosol, la mitochondrie, le glycosome. L'accepteur d'électrons final de l'ensemble du système enzymatique est le NADPH toujours dérivée de voies pentoses phosphates, ainsi que leurs équivalents réducteurs fournis dans les systèmes de détoxification par la dithiol trypanothion (T(SH2)) qui est synthétisée à partir d'un trypanothion (RTT) et la tryparedozin tioredozina (TXN) [59,68].

Trypanosoma cruzi a 5 peroxydases ayant une fonction de désintoxication: i) la peroxydase tryparédoxine cytosolique (TcCPX) et la peroxydase tryparédoxine mitochondriale (TcMPX). Les deux ont la capacité de détoxication du peroxyinitrites
la Maladie de Chagas.
endogène des macrophages (H<sub>2</sub>, O<sub>2</sub>), et de l'hydroperoxyde organique. ii) une ascorbate peroxydase hème dépendente- (TcAPX) située dans le réticulum endoplasmique (RE), conférant une résistance à H<sub>2</sub>O<sub>2</sub>. iii) des glutathiones peroxydase I et II (TcGPXI (cytosol et glicosome) et TcGPXII (RE) qui offrent une résistance contre les hydroperoxyles et les attaques des lipides. En plus de la défense résultant de ces peroxydases, T. cruzi a aussi 4 superoxyde dismutase (SOD) de fer: (TcSOD A et C), agissant dans le cytosol (TcSOD B1) et agissant dans les glicosomes (TcSOD B1-2), qui protègent le parasite contre les effets directs de O2, inhibant ainsi la formation de ONOO⁻ [60,69].

Les trypomastigotes restent dans un environnement très oxydant des phagolysosomes (détail figure 10) transitoirement et migrent 24 heures après l'infection dans le cytosol où le parasite est en phase d’amastigotes réplicatifs pendant la majorité de son cycle biologique intracellulaire [70]. Paiva et al. [71] ont démontré qu’une fois que le parasite a atteint le cytosol des macrophages, le stress oxydatif peut aussi contribuer à une augmentation de la charge parasitaire par un mécanisme qui peut faciliter l'accès des amastigotes au fer qui est essentiel pour la croissance du parasite.

Traités avec de la protoporphyrine de cobalt (COPP) (un activateur du NRF2 facteur de transcription, qui orchestre les réponses anti-oxydants), ou avec d'autres anti-oxydants, les macrophages péritonéaux de rats de laboratoire infectés par T. cruzi, ont une charge parasitaire notamment réduite. L’effet de la protection CoPP sur une infection par T. cruzi est indépendante de l'immunisation des cellules T et ne comporte pas de cellules apoptotiques infectées ou d'autres facteurs qui agissent contre la parasitémie, ce qui suggère un effet délétère des effets classiques d'adaptation antioxydantes et les réponses immunitaires innées [70,71].

La ligne de défense immunitaire contre T. cruzi.

Les récepteurs de reconnaissance des formes (REM) sont décrits comme la première ligne de défense immunitaire contre une large gamme d'agents infectieux et pathologique comme les protozoaires. Ces récepteurs sont exprimés par les cellules immunitaires et ont une fonction de reconnaissance des molécules qui sont communes aux agents pathogènes, mais qui se différencient des molécules de l'hôte. Ceci identifie une collection de motifs moléculaires spécifiques des agents infectieux (MSAI) (PAMP). Les récepteurs Toll-like (TLR), font partie des REM’s PPRs bien caractérisés et ont une capacité de détection de PAMP, qui sont localisés sur la surface des cellules, rendant ces récepteurs abondants dans les cellules et les cellules T sommatives. L’activation des TLRs
la Maladie de Chagas.

conduit à la production de cytokines pro-inflammatoires et de chimiokines qui à leur tour conduisent à la production de cellules phagocytaires dans les tissus infectés. Ces cellules sont d'une grande importance non seulement pour contrôler l'infection initiale résultant d'un nouveau corps dans l'hôte, et aussi pour la planification des réponses immunitaires adaptatives ultérieures [72,73].

Les modèles murins de laboratoire (rats) présentent un total dé 13 types de TLR et 10 ont été identifiés chez l'homme. Les TLR 1-9 sont présents chez les rats et l'homme. Les TRL 11,12 et 13 sont spécifiques des rats et le TLR 10 est spécifique de l'homme, ce qui rend cette superfamille un bon indicateur de la réponse immunitaire au parasite. Certaines formes de TLR sont sous forme d'homodimères (TLR-9 4e), tandis que d'autres sont des heterodimères (TLR 2 et 6). Après infection, ces protéines ont tendance à changer de conformation (autres que les TLR 3), et recruter TIR (Toll/IL-1R), et déclencher une cascade de signalisation dépendant de MyD88 qui résulte dans la production de cytokines pro-inflammatoires [72,74].

Infectés par T. cruzi, un grand nombre de protéines de surface de T. Cruzi, des mucines (glycoinositolphospholipids – GIPL), l'ADN et l'ARN de T.cruzi peuvent stimuler les TLR de l'homme (TLR1-10) ou du rat (TLR1-9 et TLR11-13) [75-77].

La figure 12 montre le schéma d'activation du TLR et des défenses cellulaires.
la Maladie de Chagas.

Figure 12: Processus d'adhésion de *T. Cruzi* à la cellule hôte et stimulation des TLRs. L'activation de l'hétérodimère TLR2 et 6 par des Glycosylphosphatidylinositol (GPI-mucines) du parasite peut conduire à la production de TNF dans des macrophages ou à l'inhibition de l'IL-12 dans les cellules dendritiques (flèches bleues). L'activation du TLR4 par les glycoinositolphospholipides du parasite (flèches vertes), TLR9 par des motifs CpG d'ADN du parasite (flèche violette) et TLR7 par l'ARN du parasite (flèche rose) se traduisent toutes par la production de cytokines pro-inflammatoires comme l'IL-12. Après être sorti du phagolysosome, le parasite peut activer le récepteur NOD1 cytoplasmique. Bien que ce soit le récepteur important pour contrôler l'infection, son mécanisme d'action est encore inconnu. (D’après Cardoso et al. 2016).

Les glycosylphosphatidylinositol-mucines (GPI-mucines) sont situées sur la surface du parasite et sont liés à la fonction d'échappement au système immunitaire et d'adhérence aux cellules hôtes. Dans la phase trypomastigotes, sa composition consiste en acides gras insaturés position sn-2, hétérodimères de stimulus extracellulaire TLR-2/6, résultant en la production de cytokines pro-inflammatoires comme l'IL-12 et le TNF, et les oxydes d'azote, ce qui permet la réponse immunitaire TH1(cytokines de type Th1: produire la réponse pro-inflammatoires responsables du meurtre des parasites intracellulaires et de la perpétuation de réponses auto-immunes) [78-82].

Toutefois des études in vitro et in vivo [77,83] montrent que des souris déficientes en TLR2 infectées par *T. cruzi*, développent une réponse auto-immune pro-inflammatoire plus sévère avec des niveaux élevés d'IFN-y que les souris type sauvage (WT). TLR2, peut conférer une résistance de la cellule hôte suite à l'infection par *T. cruzi*. Les autres TLRs sont présents en quantité inférieure.
la Maladie de Chagas.

Gravina et ses collaborateurs [75], ont suggéré que TLR2 assume des fonctions différentes en fonction du type de cellule hôte: production de TNF dans les macrophages et d'immunorégulateur dans les cellules dendritiques. La surface de T. cruzi comprend au plus $2 \times 10^6$ molécules de mucines, ce qui force l'action TLR2 pendant le début d'infection parasitaire, la stimulation de la régulation immunitaire cellulaire et en conséquence un retard dans la réponse immunitaire et la production d'anticorps dirigés contre T. cruzi [75, 80, 84].

D'autres aspects de l'immunité innée contre T. cruzi doivent être pris en compte. Les cellules natural killer (NK) jouent un rôle crucial au stade de initial de l'infection, car elles limitent la croissance parasitaire, et promeuvent l'immunité cellulaire [85]. Parmi les mécanismes effectués par les cellules NK, on peut distinguer un rôle lytique (destruction de cellules cibles) et la production de cytokines impliquées dans la sensibilisation des autres cellules immunitaires. Les premiers événements provoqués par une infection sont: l'hépatosplénomégalie, une lymphadénopathie locale, l'atrophie des ganglions lymphatiques mésentériques et du thymus, en plus de l'activation polyclonale des lymphocytes T et B, une hypergammaglobulinémie, suivie d'une augmentation de l'apoptose et immunosuppression de la réponse des lymphocytes [86].

Des études ont montré que l'infection des macrophages par T. cruzi peut induire la sécrétion d'interleukine 12 (IL-12) par ces cellules, ce qui conduit à une production accrue d'interféron gamma (IFN-γ) et de facteur de nécrose tumorale alpha (TNF-α) qui diminuent la parasitémie en agissant sur l'adhésion et l'internalisation du parasite. On pense également que les protozoaires activent les fonctions des cellules NK par des mécanismes indirects impliquant la stimulation et la production de cytokines par les cellules présentatrices d'antigènes. Les glycolipides sont d'une grande importance comme inducteurs de la production de cytokines (IL-12 et TNF-α), et fonctionnent comme des points d'ancrage dans l'interaction avec les cellules NK et l'induction de ces cytokines. Le rôle des cellules NK est de nature transitoire dans des hôtes infectés suggérant que son activité est strictement contrôlée [85].

**La Rupture des Cellules.**

Après le premier cycle de réplication et de la métamorphose de la forme amastigote en trypomastigotes dans la cellule hôte, T. cruzi atteint la circulation sanguine et devient la cible des protéines solubles qui interagissent avec les agents pathogènes et activent une cascade de protéases qui éliminent les micro-organismes envahissants. Le système de
la Maladie de Chagas.

défense dans le flux sanguin a trois voies: la voie classique, la voie alternative, classique
et la voie de la lectine (Figure 13A-C). Ces trois voies sont initialement différentes, mais
résultent dans la production d'une convertase C3, puis d'une convertase C5, et enfin d'un
complexe d'attaque membranaire (MAC), ce qui favorise la lyse du parasite [70]. La figure
13 résume les processus d'attaque membranaire.

Figure 13: Mécanismes d'attaque de T. cruzi dans le flux sanguin. Il y a trois voies du complément: classique, alternative
et lectine. (A) Dans la voie classique, les anticorps liés aux antigènes d'agents pathogènes interagissent avec la protéine
du complément C1, qui clive C2 et C4 pour générer C2a et C4b; Ces molécules se lient à la surface de l'agent
pathogène pour former la C3 convertase C4b2a. (B) Dans la voie de la lectine, la MBL se lie au mannane, à la ficoline
ou à des molécules glycosylées, présents sur la surface de l'agent pathogène, et les protéases à cystéine liée à ces
molécules clivent les convertases C2, C3 et C4 générant le C4b2a. (C) Dans la voie alternative, C3b clive spontanément
d'autres C3b, et la forme clivée interagit avec le facteur Bb qui provient d'un clivage du facteur B par le facteur D. le
facteur Bb se lie à convertase C3b, ce qui crée une convertase C3bBb qui se transforme en convertase C3. La
convertase C3, effectue une interaction complète avec C3b pour former une convertase C5, qui se clive en C5b. (D) C5b
interagit avec C6-9 pour former le MAC, conduisant à la lyse des agents pathogènes. (E) Réponse de T. cruzi, pour
eviter la lyse, en utilisant la calréticuline et gp58 / 68 (Gp58), qui bloquent les étapes Initiales de la voie classique et de
la voie lectine. En plus CRIT (l'inhibition du récepteur trispanning), T-DAF (trypomastigotes décroissance-facteur
d'accélération) et CRP (protéine régulatrice du complément), perturbent l'assemblage ddu bloc C3 convertase. Ag :
antigène; Carb: hydrate de carbone. CALRE: calréticuline. ( D'apres Cardoso et al. 2016).

T. cruzi cible initialement la voie de la lectine. Elle est activée par la liaison des
carbohydrates à la surface du parasite aux cellules sanguines. Cette voie de la défense
est responsable de plus de 70% des cas de rupture du parasite cours de l'infection.
Lorsque l'infection progresse, les anticorps anti T. cruzi sont produits et se lient à la
surface du parasite, en essayant d'interagir avec les convertases C1, ce qui provoque
l'inactivation du parasite [87,88].

Pour échapper à cette action d'anticorps, T. cruzi possède un grand nombre de
protéines de surface qui bloquent les diverses formes de défense du complément. T. cruzi
compromet la voie lectine par l'intermédiaire de la calréticuline, une protéine de surface de
45 kDa [89]. Cette protéine se lie aux carbohydrates de surface de T. cruzi, ce qui
empêche la convertase C1, d'effectuer l'inhibition de la voie classique. La connexion de la
la Maladie de Chagas.
surface du parasite à la cellule hôte peut aussi être évitée par la L-ficoline, qui permet la conversion de C4 à C4b, mais la calréticuline de T. cruzi évite la formation de C1 ce qui influe sur la production de C4, et donc diminue l'efficacité de la voie classique et de la voie de la lectine [89-92]. Une autre façon d'éviter la voie classique et la voie de la lactine vient des protéines complémentaires de régulation (PCR) appelées GP160, qui sont des protéines situées à la surface de T. cruzi et qui se lient à convertase C4b et C3b, et empêchent que la convertase C3 se connecte à C3b et C4b [93]. Beucher et Norris décrivent plusieurs étapes concernant l'action de la PCR, de différentes natures (chimiothérapeutiques, défenses naturelles de), mais ces études nécessitent une validation plus expérimentale plus poussée [93,94].

**Les cellules T-CD+ 4 et 8**

Les lymphocytes T-CD4⁺ sont principalement impliqués dans l'activation et la régulation des autres cellules et sont donc appelés agents auxiliaires [95]. Ce groupe de lymphocytes est divisé en sous-populations fonctionnellement distinctes, en raison de la diversité des cytokines produites: Th1 et Th2 (sous-groupe, qui agit par l'intermédiaire des cytokines anti-inflammatoires IL-4 et l'IL-10, et participe principalement à l'induction de la réponse immunitaire, considéré comme essentiel dans le processus infectieux généré par T. cruzi) (voir Figure 14) [96].

Les T-CD8+ sont un autre groupe important de cellules qui sont impliquées dans la production de cytokines et de l'activité cytolytique, contrôles de la réplication parasite in vivo et de l'hôte résistant [97]. L'action de défense par CD8⁺ et CD4⁺ se produit par médiation cellulaire, ce qui induit la production d’ozone nitrique (NO) par les macrophages, un élément important pour déclencher la réponse Th1 contre T. cruzi, et contrôler la réplication par la sécrétion de cytokines (TNF-α, IL-1, IL-l’12) [96].

En plus des cellules T-CD, les cytokines sont des médiateurs importants de la durée et de l'intensité de la réponse immunitaire. Elles sont produites par les leucocytes et d'autres types de cellules qui agissent comme médiateurs de produits chimiques. En fait, la résistance à l'infection de la cellule hôte au cours de la phase aiguë est directement liée à l'activation des macrophages par des cytokines [98,99].
la Maladie de Chagas.

Figure 14: Les mécanismes associés au contrôle de la charge parasitaire sont surlignés en vert. Les agents associés à la modulation parasitaire du système immunitaire de l'hôte et / ou à une augmentation de la charge parasitaire sont mis en évidence en rouge (D’après Perez et al. 2002).


la Maladie de Chagas.


26. Dias LC, Dessoy MA, Silva JJN, Thiemann OH, Oliva G, Andricopulo AD Quimioterapia
la Maladie de Chagas.


la Maladie de Chagas.


50. Barreto-de-Albuquerque J, Silva-dos-Santos D, Pérez AR, Berbert LR, de Santana-


la Maladie de Chagas.


77. Oliveira A-C, Peixoto JR, de Arruda LB, Campos MA, Gazzinelli RT, Golenbock DT,
38


la Maladie de Chagas.


la Maladie de Chagas.

CHAPITRE II
la Maladie de Chagas.

**Traitements existants contre la maladie de chagas.**

L'arsenal médicamenteux contre la maladie de Chagas est actuellement faible. Les médicaments utilisés dans le traitement de la maladie de Chagas sont essentiellement deux nitroaromatiques hétérocycliques. Le premier est le benznidazole (BZ) (2-nitroimidazole (N-benzyl-2-nitroimidazole) acétamide), introduit sur le marché depuis plus de trois décennies [1]. Il est produit par le laboratoire de l'État de Pernambuco – Brésil) et commercialisé sous le nom Rochagan au Brésil, et Radanil en Argentine [1]. Le second est le nifurtimox (NFX) (3-méthyl-4- (5'-amino-nitrofurfurilideno) -tétrahydro-4H-1,4-thiazine-l, 1-dioxyde) . Il est commercialisé par Bayer, mais a maintenant été retiré du marché au Brésil) [2]. La Figure 15 montre les structures 2D des deux composés.

![Figure 15: 2D structure du nifurtimox (à gauche) et du Benznidazole (à droite). (Générées par Marvin Beans Version 16.10.3.0)](image)

Deux hypothèses peuvent expliquer le mécanisme d'action du Nifurtimox. L’une implique la formation d’un nitroanion qui réagirait avec les acides nucléiques du parasite provoquant une chute significative de l’acide désoxyribonucléique (DNA) dans un mécanisme similaire à celui proposé pour les autres nitrofuranes antibactériens [3]. L’autre implique la production de superoxydes et, par conséquent, le peroxyde d’hydrogène (les deux sont très toxiques pour le parasite) et l’inhibition de la trypanothione-réductase, qui est une enzyme du parasite spécifique pour la défense contre les radicaux libres oxygénés [3]. L’absence de ces enzymes conduit à l’accumulation de peroxyde d’hydrogène à des niveaux cytotoxiques, entraînant la mort du parasite. Les effets secondaires les plus fréquents de traitement par NFX sont l’anorexie, la perte de poids, des changements mentaux, l'excitabilité ou la somnolence et des troubles gastro-intestinaux tels que nausées, vomissements, crampes intestinales et diarrhées [4].

Le mécanisme d'action du Benznidazole semble être du à sa liaison aux composants de la cellule: protéines, lipides ou de l'ADN du parasite [4]. Les deux
la Maladie de Chagas.

complications les plus graves induites par la prise de BZ sont (i) une neutropénie (diminution du nombre de polynucléaires neutrophiles circulants inférieure à 1500 / mm³) suivie d'une agranulocytose, et (ii) fièvre, septicémie et purpura thrombocytopénique, caractérisé par une réduction des plaquettes, des pétéchies, des cloques hémorragiques et des saignements des muqueuses [4].

NFX et BZ sont efficaces pour réduire la durée et la gravité de la maladie de Chagas aiguë et congénitale, avec un taux de guérison d'approximativement 50% des patients traités [2]. Cependant, avec BZ, on observe la guérison de 76% des patients en phase aiguë mais seulement 8% des patients en phase chronique, et chez les enfants, le taux de guérison est de seulement 55,8%. Il faut noter que des effets indésirables graves ont été rapportés lors de l'utilisation clinique par les médecins ou par les patients : perte de poids, nausées et vomissements, excitabilité nerveuse, insomnie, convulsions, étourdissements et manque de mémoire sont fréquemment observés [2].

Finalement, Il a été constaté que certaines souches de T. cruzi, y compris certaines présentes dans un environnement sauvage sans contact avec les patients traités possèdent une résistance à ces médicaments. Un autre point important est la dépendance de l'efficacité de BZ au système immunitaire de l'hôte : L'efficacité du traitement par le BZ est réduite chez les patients présentant une déficience du système immunitaire [5].

En conséquence, plusieurs études portant sur le traitement de la maladie de Chagas ont été faites jusqu'à ce que, en 1998, l'Organisation mondiale de la santé (WHO) réunisse un groupe de quinze experts d'Amérique latine et détaille les exigences de base pour qu'un composé puisse être un candidat pour le développement d'un médicament contre la maladie de Chagas: (i) guérison parasitologique en cas de maladie aiguë et chronique; (ii) l'efficacité de protocoles courts (seule dose pour la phase aiguë) ou longs (plusieurs doses pour la phase chronique); (iii) un faible coût permettant de cibler la population à risque; (iv) l'absence d'effets secondaires ou tératogènes; (v) une administration possible en ambulatoire ou par soins à domicile et (vi) la non induction de résistance [4].
Solutions thérapeutiques en cours d'exploration pour le traitement de la maladie de Chagas.

Peptides antimicrobiens - Gramicidine A

Les peptides antimicrobiens (PAMs) sont largement étudiés car ils ont la capacité de lyser la membrane cellulaire parasitaire par la formation de pores dans la membrane ce qui se traduit généralement par la mort cellulaire ou l'induction de l'apoptose [6,7]. La Gramicidine A est un exemple de PAM antibiotique de 15 acides aminés provenant de Bacillus brevis. Dans la membrane cellulaire, ce peptide forme des canaux qui permettent le passage spécifique des cations monovalents. Ces canaux correspondent à des dimères de deux structures hélicoïdales qui peuvent adopter quatre conformations (voir Figure 16): (a) l’interaction monomère-monomère par la partie N-terminale; (b) Interaction carboxyle-carboxyle pour la formation de dimères; (c) double hélice d’antiparallèle, et (d) double hélice parallèle, formée en contactant la bicouche lipidique du parasite [8].
la Maladie de Chagas.

Des études ont démontré que l'action de la Gramicidine dépend du nombre de canaux ioniques monovalents formés, et chez le parasite, ce nombre est beaucoup plus élevé que dans les cellules de l'hôte [9,10]. Cependant, la dose administrée est d'importance cruciale pour l'effet antiparasitaire. Une utilisation excessive augmente le niveau de toxicité, et provoque des effets secondaires graves. Il demeure que ce type d'approche pourrait déboucher en termes d'innovation thérapeutique, par la recherche de peptides naturels ou de synthèse ayant un potentiel contre la maladie de Chagas [10].

**Peptides de venin - Mélittine**

La Mélittine est un composant toxique majeur du venin de *Apis mellifera*. Elle a un poids moléculaire de 2846,46 Daltons et comprend 26 acides aminés avec une séquence hydrophile (Lys-Arg-Lys-Arg) à proximité de région C-terminale. Son action contre le parasite se fait par la formation de pores dans la membrane cellulaire, ce qui conduit à la lyse du parasite [11]. Dans le cas de *T. cruzi* ce peptide a une action forte contre les formes épimastigotes et trypomastigotes, et de manière non-toxique pour les cellules hôtes. Ce peptide peut aussi provoquer des changements morphologiques du parasite dans leur forme amastigote, ce qui correspond à l'induction de différentes voies de la mort cellulaire, également sans conséquences graves pour l'hôte [11,12].

La mort cellulaire programmée (MCP) est un processus génétique régulateur de l'homéostasie de divers organismes. Elle dépend des conditions environnementales et morphologiques et est divisée en trois types distincts: Apoptose (I-MCP), la nécrose programmée (II-MCP) et autophagie finale (III-MCP) [13]. L'apoptose est caractérisée par une série de facteurs tels que le rétrécissement cytoplasmique, la condensation de la chromatine, la fragmentation de l'ADN chromosomique, un gonflement des mitochondries et autres. L'autophagie est un processus complexe qui implique plus de 30 protéines ATG (protéines liées au processus d'autophagie) avec un remodelage structurel. Lors de l'emploi de la mélittine, les parasites exposés possèdent plusieurs de ces modifications [11]. La figure 17 montre en détail les modifications du traitement mélittine dans la forme épimastigote.
la Maladie de Chagas.

Figure 17: L’analyse ultrastructurale de formes épimastigotes traitées par la melittine. (A, D) Les parasites non traités affichent le corps typique allongé cellulaire et une morphologie nucléaire normale (N), un kinétoplaste (k), mitochondrie (M) et poche flagellaire (FP). (B, C) Les parasites traités présentent une conformation enflée et anormale du corps cellulaire (astérisques) et une morphologie flagellaire modifiée avec des fissures, des structures gonflées, et une apparence brisée (flèches). L’analyse ultrastructurale des parasites traités révélé des altérations mitochondriales, tels que le gonflement (E, F) et des modifications des membranes intérieures avec la formation de figures concentriques (flèches noires en F). Dans les kinétoplastes des modifications sont aussi observés (E). Les caractéristiques mises en évidence notamment la présence de profils de réticulum endoplasmique (pointe de flèche G) suggèrent des autophagosomes (AP) et sont indicatives de la mort par autophagie (G, H, encadré). Des réeservosomes (R) avec des contenus désorganisés sont également observés (étoiles à I). Bars: A-C: 1 µM; D-I: 0,5 µM. (D’après Adade et al. 2013).

Sur les formes trypomastigotes l’effet de la mélitine est une enflure mitochondriale et une diminution du corps de la cellule. L’effet le plus important se situe dans les régions du kADN et le noyau, avec de grands changements dans les groupements filamentieux et

45
la Maladie de Chagas.


la Maladie de Chagas.

**Dérivés de phtalazine.**

Les 1,4-Bis(alkylamino)benzo[g]phthalazines, (Figure 19:B) ont la capacité de complexer les cations métalliques (Cu, Zn, Fe et Mn) dans des cellules infectées par des parasites, ce qui conduit à l'interruption de l'infection et la mort du parasite. Cette caractéristique est due à la présence de groupements pyridazine qui interagissent avec des cations métalliques [14].

Figure 19: A) Structure du benzo [g] phtalazine. B) Structure de 1,4-bis (alkyl amino) benzo [g] phtalazine. Les régions R1 et R2 susceptibles de modification. (D'après Sánchez-Moreno et al. 2012).

Les dérivés de la 1,4-Bis(alkylamino)benzo[g]phthalazines, avec des chaînes latérales de type alkylamine, ou avec des unités hétérocycliques (comme la pyridine, pyrazole ou imidazole) (Figure 20) ont soulevé un intérêt récent dans le cadre le la maladie de Chagas. Dans des tests in vivo chez des souris infectées par T. cruzi, on a
la Maladie de Chagas.

observé une faible toxicité contre les cellules murines, mais une forte toxicité pour T. Cruzi dans les deux stades de l'infection (aiguë et chronique) [14,15]. Les études d'histopathologie ont montré que les composés avec des cycles imidazole ont une toxicité pour les cellules murines plus faible que les composés avec des pyrazoles tout en ayant une activité anti-Cruzi du même ordre.

Figure 20: Dérivés de phthalazine ayant une activité anti-parasitaire potentielle (D’après Sánchez-Moreno et al. 2011).

La cible de ces composés semble être l'enzyme superoxyde dismutase fer (Fe-SOD), qui ne se trouve que chez les parasites et a un rôle clé dans la défense contre le stress oxydatif causé par réponse auto-immune de l'hôte à l’infection. L'interaction avec les ions présents dans la région du site actif pourrait déstabiliser l'action antioxydante de l'enzyme. Les dérivés mono alkylamino substitué de type II sont des inhibiteurs plus puissants de Fe-SOD que les analogues disubstitués (Figure 19B:2). Parmi les composés représentés (figure 20), ceux du cadre 2 et 4 ont des valeurs expérimentale d'inhibition de la Fe-SOD remarquables, 5,5 et 5,8 fois plus élevées que celle du contrôle BZN. Enfin, fait intéressant, leur capacité d'inhibition de la CuZn-SOD humaine est négligeable. Au niveau cellulaire, l'effet de ces composés induit des changements (Figure 21), parmi lesquels : i) de modifier la pochee du flagelle, qui rend plus difficile la mobilité du parasite - composé 2.
la Maladie de Chagas.

ii) la modification de la membrane nucléaire avec des unités de séparation - composé 3.

iii) le gonflement des mitochondries - composé 4 [14-16].

Pour la lutte contre la maladie de Chagas, les résultats obtenus en phase chronique de l'infection parasitaire sont particulièrement prometteurs, car c'est à ce stade que la maladie devient très résistante aux thérapies usuelles [15]. Enfin, ce type de composé réduit aussi l'activité parasitaire d'autres espèces comme L. braziliensis et donovani impliqués dans la leishmaniose.

Figure 21: Transformations Ultrastructurales observées par microscopie électronique à transmission pour les épimastigotes de T. cruzi non traitées (contrôle) ou traités avec des composés 1-4 de la Figure 19. noyau (N), réservosomes (R), vacuoles (V) mitochondrie (M), glycosomes (G), microtubules cytosquelette (MI), de grandes poches flagellaires (FB), la réduction de la taille du parasite (flèche), nucléaire membrane modifiée (MN) (D’après Sánchez- Moreno, et al. 2012b).
Terpenes - Acide bétulinique

L'acide bétulinique est un triterpène pentacyclique abondant dans le règne végétal. Il peut être isolé à partir de plusieurs espèces de plantes ou obtenu à partir de son précurseur métabolique, la bétuline. L'acide bétulinique et ses dérivés possèdent une activité anti-VIH, anti-bactérienne, anthelminthique (anti-parasitaire, classe de médicament), anti-inflammatoire et une activité cytotoxique puissante contre un large panel de lignées cellulaires tumorales [17-19].

Des études on montré que l'activité anti-T. Cruzi de l'acide bétulinique et d'esters dérivés inhibe la prolifération de la forme épimastigote. Compte tenu de ces résultats, l'acide bétulinique est considéré comme une tête de série pour la conception et la synthèse des agents antiprotozoaires. Les modifications chimiques du groupe carboxyle peuvent résulter dans des dérivés ayant une activité antiprotozoaire améliorée en comparaison à l'acide bétulinique [20-22]. La figure 22 représente la principale région AB, et les régions R1 et R2, qui peut être modifiées.

Figure 22: L'acide bétulinique et les régions modifiables. ([D'après Domínguez-Carmona et al. 2010]).

A partir de l'idée d'une changement structural dans AB, un nouveau groupe de 8 produits (figure 23) a été récemment été généré pour identifier leurs implications dans l'action anti-chagas [17-19].
la Maladie de Chagas.

Cette recherche a permis de comprendre que l'inclusion de groupes morpholine, thiomorpholine et les groupes 4-methylphenylpiperidine potentialisent les actions antiparasitaire. Ils induisent des changements dans la morphologie du parasite. Par exemple, le traitement par le composé BA5 provoque une diminution du flagelle, la perte de l'intégrité de la membrane plasmique et une grande déformation du corps, accompagnée de la formation de vacuoles anormales dans le cytoplasme, ainsi que l'expansion de l'appareil de Golgi. Ces changements sont liés à la mort du parasite, due à une nécrose [18,20]. En plus de la mort du parasite, ces substances semi-synthétiques, provoquent aussi l'empêchement du développement du parasite et l'invasion de la cellule, des étapes fondamentales de la propagation de la maladie de Chagas. On observe donc une certaine similitude d'action par rapport au benznidazole. Un effet imprévu de ces travaux a été une augmentation des recherches sur des produits naturels, en particulier ceux issus de l'Amazonie brésilienne [21].
La recherche de nouvelles cibles thérapeutiques contre la maladie de Chagas.

Pour une pathologie d'intérêt, l'identification de nouvelles cibles biologiques à l'aide d'outils bio-informatiques bénéficie désormais d'une grande quantité d'information sous forme numérique [23]. Ces données peuvent être issues de programmes de génomique, transcriptomique, protéomique, métabolomique, pharmacogénomique, parmi d'autres. Cette information doit être considérée par les chercheurs pour identifier les cibles les plus pertinentes pour la conception de médicament, et particulièrement anticiper autant que possible l'absence d'effets secondaires. Une approche génomique fonctionnelle combinant la bioinformatique, la biochimie et de biologie moléculaire et cellulaire peut permettre l'identification de nouvelles cibles potentielles [24]. Récemment un certain nombre de nouvelles cibles a été proposé dans le cadre de la lutte contre la maladie de Chagas et ce chapitre les décrit brièvement.

**Tiol Transferase (TC52)**

L'enzyme: thiol transferase (TC52) présente une similitude avec glutathion-S-transférase, une enzyme présente dans différents organismes y compris *H. sapiens*. Les Trypanosomatides n'ont pas un système de défense redox basé sur le glutathion (GSH) ou les flavoproteines-NAD. Le rôle de la TC52 est de contribuer à la détoxification de l'environnement cellulaire. TC52 est présente seulement dans les formes: épimastigote et amastigote, formes intracellulaires. Elle est impliquée dans la réponse au stress oxydant engendré par la réponse immunitaire de la cellule hôte [25-28]. La TC52 est présente sous forme d'un monomère, alors que le glutathion des cellules hôtes existe sous forme de dimère, ce qui est dû à une duplication de gènes dans l'évolution de espèces ancestrales [28].

La TC52 est associée à la virulence du parasite, et contribue à la non fonctionnalisation des défenses de la cellule hôte basées sur la production de cytokines et de NO qui sont préjudiciables aux parasites. Les études montrent qu'un vaccin à ADN ciblant TC52, peut stimuler chez la souris la production d'anticorps spécifiques de cette protéine, ce qui diminue les défenses du parasite contre l'attaque oxydante par la cellule hôte. En fait, on a pu observer effectivement une diminution des dommages cardiaques provoqués par le parasite [29,30]. Cependant, bien que ce type d’approche ait provoqué un grand espoir, aucune application chez l’homme n’a pu être menée au bout à ce jour.
Cruzipaïne 1 et 2

Les cruzipaines (superfamille englobant les cruzipaines, les falcipaines, etc) sont le groupe le plus important de cystéine-protéases trouvée dans T. cruzi. Ces enzymes sont liées à de multiples fonctions comme l'infection de la cellule hôte, la réplication cellulaire, la différenciation des cellules T. cruzi, le métabolisme et la protection contre la réponse immunitaire [31]. La présence importante de ce groupe dans différentes fonctions du parasite en fait une cible intéressante pour le développement de nouveaux thérapeutiques contre la maladie de Chagas depuis le début des années 1990 [32].

Dans la forme épimastigote, le nombre de cruzaines est 10 fois plus élevé que dans les autres formes, et sont retenus dans un organite pré-lysosomal, le réservosome. Dans la forme amastigote ces cruzaines sont présentes dans la même région, qui est connectée à la face externe de la membrane plasmique du T. cruzi, en contact direct avec le cytoplasme de la cellule infectée [31].

La partie mature de la Cruzipaïne (sans le domaine DUF) se compose de 345 acides aminés, et la région centrale comprend 215 acides aminés (22,5 kDA), avec un site actif impliquant une cystéine, une histidine et une asparaginique. Un quatrième résidu glutamine est impliqué dans maintien des composés dans le site catalytique [33,34]. La Cruzipaïne se compose de 4 sous-sites principaux (S3, S2, S1, S1') [33,34]. (voir Figure 24), et actuellement, il est proposé que les activités enzymatiques des isoformes cruzipaine pourraient être modulées par des changements conformationnels dans la région du site de liaison [35].

Figure 24 : Topologie du site catalytique de cruzipaine. (D'après Kerr et al. 2009)

A cause des nombreuses fonctions effectuées par les cruzaines, et la présence de cette famille de protéine dans toutes les formes de Trypanosoma cruzi, son potentiel en
la Maladie de Chagas.

tant que cible thérapeutique de nouveaux médicaments contre la maladie de Chagas semble évident. Très récemment, Sbaraglini et ses collègues [36] ont proposé une combinaison de deux inhibiteurs de la cruzipaine (Benidipina – un bloquant du canal calcium et la Clofazimine – utilisée pour le traitement de la lèpre ), et ont montré qu'elle est en mesure de réduire la charge parasitaire dans les tissus infectés.

Actuellement cependant, la drogue candidate de référence contre les cruzaine est K11777 (N-méthyl-pipérazine-Phe-homoPhe-vinylsulfone-phényle), un inhibiteur sélectif des cystéine-protéases, actuellement en phase pré-clinique de recherche contre la maladie de Chagas. Les tests animaux (souris) indiquent qu'il semble sans effet secondaire pour les hôtes infectés [37,38].

**Trans-sialidase de Trypanosoma cruzi (TS)**

La TS est un membre de la famille des hydrolases glycosidiques (de GH333) qui est classé comme un exo-α-sialidase (CE 3.2.1.18). La TS de T.cuzi catalyse le transfert d'acide α-2,3-β-galactose sialique terminal de l'hôte à d'autres unités β-galactosyle de mucines de la surface du parasite. Elle a une structure composée de deux parties principales: i) un domaine catalytique constitué de 6 feuillets β reliés par une longue α-hélice ii) un domaine C-terminal antigénique de la lectine de type 12 [37]. Des séquences répétées d'acides aminés composent l'antigène de la trans-sialidase. Le site actif de l'enzyme est caractérisé par la conservation d'une triade d'arginines (ARG 35, ARG245 et ARG314), du glutamate (GLU357) et de l'acide aspartique (ASP59). La poche hydrophobe, conservée dans toutes les sialidase, contient un groupe N-acétyle de l'acide sialique, qui est protégé par la chaîne latérale des acides aminés VAL95, LEU176 et TRP120 [39]. Les TS sont des enzymes situées dans les membranes cellulaires et sont essentielles pour la survie du parasite. Il faut savoir que les trypanosomatides sont incapables d'effectuer la biosynthèse de l'acide sialique. L'expression de cette enzyme est considérée comme significative dans la forme trypomastigote [39]. Cette enzyme est également liée à l'évitement de la réponse auto-immune de l'hôte à l'infection, ce qui permet la survie des trypomastigotes.

L'inhibition de cette protéine dans la forme infectieuse de T. cruzi diminue le nombre de défenses parasitaires contre les attaques externes, promues par la réponse immunitaire de l'hôte infecté, ce qui permet l'interruption de l'infection par T. cruzi et même l'avancement de la maladie de Chagas. Des études ont montré que la dansyl-NeuNAcFP (5-acetamido-2-(4-N-5-dimethylaminonaphthalene-1-sulfonyl-2-difluoromethylphenyl)-3,5-
la Maladie de Chagas.
dideoxy-D-glycero-α-D-galacto-2-nonulopyranosonic acid) est capable d'inhiber TS irréversiblement, d'une façon dose-dépendante et d'entraîner la diminution de l'invasion des cellules hôtes par T. cruzi. En outre, l'incubation de T. cruzi forme trypomastigotes en présence de 1 mM et 10 mM diminue de manière significative la capacité d'infecter des cellules (34% et 90%, respectivement) [40].

**Glycéraldéhyde 3-phosphate déshydrogénase (GAPDH)**

La glycéraldéhyde 3-phosphate déshydrogénase (GAPDH, EC 1.2.1.12) de T. cruzi fait partie d'une étape de la glycolyse: la phosphorylation oxydative de D-glycéraldéhyde-3-phosphate (G3P) en 1,3-biphosphoglycérate (1 3-BPG) avec une réduction parallèle de nicotinamide adénine dinucléotide (NAD⁺) en NADH. La GAPDH est une enzyme clé dans la voie de la glycolyse de *T. cruzi*, qui est la principale source d'énergie du parasite dans la forme trypomastigotes. Bloquer cette enzyme cause la mort du parasite en quelques minutes. Même s'il y a présence de GAPDH dans *H.sapiens*, la grande distance évolutive comparée à trypanosomatides ne se traduit pas par un effet secondaire à l'hôte [41].

Des études récentes ont démontré l'efficacité d'un composé à ruthénium (cis - [Ru(NO)(bpy)2 SO 3] PF 6) pour inhiber la GAPDH de T. cruzi. Ce composé a montré une activité élevée (IC50 89 ± 8) chez des modèles des souris de laboratoire, et il présente une activité 10 fois plus élevée sur formes tripanomastigotes que le médicament de contrôle recommandé par l'OMS (BZ). L'activité de ce composé est également observée pour la forme épimastigote en réduisant la prolifération. Ce composé est aussi actif contre les amastigotes, à l'intérieur de la cellule. Sur des souris infectées, on a pu parvenir à une diminution de 75% du nombre de parasites. La grande capacité d'inhibition de ce composé pour les différentes formes de T. cruzi a accrédité l'hypothèse qu'il est un bon candidat pour les deux phases de la maladie de Chagas (aiguë et chronique) [42]. Cependant, aucune étude plus récente ne semble confirmer cette hypothèse même si ce type de composé est aussi en cours d'évaluation dans d'autres cadres comme celui de la Leishmaniose [43].

**Sterol 14α-demétilase (CYP51)**

La superfamille des monooxigénases contient plus de 12 000 membres, et est présente dans tous les royaumes du vivant. Cette superfamille peut être décrite en deux grandes familles, l'une qui dégrade des xénobiotiques et une autre qui est requise pour la biosynthèse de composés endogènes [44-46]. En raison de sa large distribution dans
la Maladie de Chagas.

différents organismes, une forme se distingue par son caractère unique au milieu biologique, la stérol 14α-déméthylase (CYP51). Cette forme est essentielle à la biosynthèse des stérols qui constituent la membrane [47,48]. Les études sur son inhibition ont commencé dans les maladies simples telles que l'infection par Candida albicans, par le VIH, puis le cancer et atteint désormais les protozoaires comme Leishmania et trypanosomes [48].

CYP51 de T. cruzi est une monooxygénase dont la fonction est d'éliminer le groupe méthyle de l'14α-eburicol, empêchant la réaction de dégradation des stérols de la membrane du parasite [49,50]. Pour le maintien de l'intégrité de sa membrane plasmique, T. cruzi dispose d'une présence massive de CYP51, ce qui le rend sensible aux attaques de inhibiteurs du stérol. Cette dépendance rend CYP51 de T. cruzi une cible attrayante pour l'identification de nouvelles drogues pour lutter contre la maladie de Chagas. De plus, dans le cadre de la lutte contre la forme trypomastigote, ce type de stratégie ne peut en principe pas générer de dommages colatéraux aux hôtes vertébrés, car leur dépendance au stérols peut être réduite par d'autres voies de rechange [49].

Des études récentes ont démontré la capacité des dérivés aminopyridyle-4 pour inhiber CYP51 dans les différentes formes de T. cruzi. Ces composés qui possèdent une grande biodisponibilité et une faible demi-vie, n'ont pas d'incidence sur la fonction de la CYP51 humaine. Cette activité anti T. cruzi est due à l'épuisement des niveaux de stérols de T. cruzi, dans un temps moyen de 3 à 4 jours de traitement dans des modèles murins infectés [51].

**Conclusion**

A ce jour, la lutte contre la maladie de Chagas s'est diversifiée. Les travaux pour développer des médicaments innovants explorent des pistes très diverses qui vont de la recherche de composés naturels, ou l'exploitation de peptides antimicrobiens à des petits composés chimiques ciblant des enzymes spécifiques. Cependant aucune stratégie ne semble à ce jour capable de mettre sur le marché l'un composé satisfaisant à tous les critères énoncés par le WHO en 1998. Il demeure important de poursuivre l'identification de nouvelles cibles, ce qui a été l'objet de mon travail de thèse.


la Maladie de Chagas.


la Maladie de Chagas.


la Maladie de Chagas.

therapeutic DNA vaccines against Trypanosoma cruzi in mice. FEMS Immunology & Medical Microbiology 50: 333-341. (2007)


44. Lepesheva GI, Waterman MR Sterol 14α-demethylase cytochrome P450 (CYP51), a P450 in all biological kingdoms. Biochimica et Biophysica Acta (BBA) - General Subjects 1770: 467-477. (2007)


49. Doyle PS, Chen C-K, Johnston JB, Hopkins SD, Leung SSF, Jacobson MP, Engel JC,


Aims of the Thesis

CHAPITRE III
Aims of the Thesis

According to the World Health Organization (WHO), 21 Latin American countries are endemic for Chagas disease, which today affects 10 million people. Chagas’ disease, caused by the protozoan Trypanosoma cruzi (T. cruzi), is a parasitic illness endemic mostly in Latin America. This parasite has a digenetic life cycle with four major life stages: epimastigotes and metacyclic trypomastigotes in the insect vectors and trypomastigotes and amastigotes in the mammalian hosts. So far there is no vaccine or effective chemotherapy treatment considering that the drugs commonly used to treat human infection, nifurtimox (3-methyl-4-(5'-nitrofurfurilidenoamino)-tetrahydro-4H-1,4-thiazine-1,1-dioxide) and benznidazol (N-benzyl-2-nitro-imidazol-1-acetamide), have high toxicity and do not prevent the progression of the chronic form of Chagas disease. Among the side effects of nifurtimox are anorexia, weight loss, psychic excitability or drowsiness and digestive symptoms. In contrast, the treatment with benznidazole causes hypersensitivity dermatitis or digestive intolerance and may lead to bone marrow depression, liver toxicity as well as polyneuritis.

Before cell invasion, T. cruzi must access and form stable adhesions to the host cell surface. Some T. cruzi surface proteins are able to bind to components of extracellular matrix or to host cell surface receptors, and others display hydrolytic activities against components of the extracellular matrix. As a result, plasma membrane proteins are essential to safeguard cell integrity and for cell/cell communication, membrane transport, as well as transmembrane signaling processes. But many other proteins can also contribute to the defense against the pathogen [1].

Irrespective of the nature of the disease, there is a need for an approach that enables the identification of enzymatic functions specific to a particular pathogen and/or drugs capable of inactivating a parasite specific enzyme without affecting humans and overall the right strategy is to seek out proteins vital for the parasites and not for the host.

A number of screening campaigns, and the development of new and improved in vitro and in vivo assays, has led to advances in the field of drug discovery. In particular, high throughput screening (HTS) has contributed to an increase of the number of hits and leads issued from Chagas disease drug discovery. With automated screening, hundreds of thousands of molecules were tested for the first time against the T. cruzi, either in viability assays suitable to automation or against specific isolated structures (i.e. target-based
Aims of the Thesis

The largest screening campaigns include the screen of approximately 300,000 compounds by the Broad Institute [2] and, recently, 1.8 million compounds from the GSK collection [3]. In both cases, the primary screening was performed using the beta-galactosidase T. cruzi assay, which has a colorimetric readout and provides an indirect measure of number viable amastigotes in host cells. These screening have resulted in the selection of a large proportion of CYP51 (Sterol 14α-Demethylase Cytochrome P450) inhibitors, however, recent clinical studies have shown these inhibitors have unsustained efficacy in patients [4].

The knowledge of the molecular target and mechanism of action of a compound can greatly aid the development of optimized molecules and facilitate further compound development, especially regarding chemical optimization of the compound. Forward chemical genetics approaches and whole-genome sequencing, proteomic, and metabolomics studies of drug resistant strains are strategies that may help to identify targets of compounds emerging from phenotypic screening campaigns [5,6]. Along, with these approaches, several proteins have been proposed as new biotargets against Chagas disease and 3D models were built based on homology modeling followed by molecular dynamics simulations. These T. cruzi biotargets include NADH-dependent fumarate reductase, ribosomal P0 antigenic protein, silent-information regulator 2 proteins or sirtuins, a putrescine–cadaverine permease or tubulin [7], among others. While these studies provide useful information on specific targets, a large-scale search of promising proteins is desirable.

In this context, the analogous enzyme pipeline (AnEnPi), which compares genomic datasets for analogous enzymes by clustering the primary structures of enzymes with the same described activity, is a tool able to identify enzymatic activities that may be potentially useful for the design of new therapeutic targets. AnEnPi searches for enzymes specific to a parasite and enzymes that are analogous for the same enzymatic function, but structurally distinct in the parasite and its human host. Using AnEnPi with a Blastp similarity raw score of 120 as cut-off, Capriles et al. identified 41 protein sequences classified as analogous or specific for T. cruzi compared to H. sapiens on the basis of similarity detection by the local alignment program BLAST, and then used automatic 3D structure modeling to infer differences between the host and parasite protein structures [8].

The goal of my thesis is to revisit these 41 protein candidates both in terms of specificity and 3D modeling by using a more complex strategy and more accurate
Aims of the Thesis

techniques. Indeed, it is known that profile-profile alignments are more sensitive than the local alignment program PSIBLAST for similarity, and automatic 3D modeling methods lead to significant structural errors for sequence identities < 80%. The workflow for *T. cruzi* protein modeling we developed consists in the following steps: sequence clustering, prediction of cell localization and enzyme classification (EC) number and analysis of metabolic pathways, homology search in *H. sapiens* and other organisms, 3D template identification, model generation and refinement by molecular dynamics (MD) simulations. Details of the model generation procedure used are described in the Chapter IV Materials and Methods and the full workflow is presented in the following chapter.

Overall, we find that it is possible to cluster the 41 protein sequences identified as analogous or specific for *T. cruzi* versus *H. sapiens* into 7 groups, and based on the current Protein Data Bank (PDB) structures, it is possible to provide structures for *T. cruzi* sequences with clear homologs or analogs in *H. sapiens* and likely associated with trypanothione reductase, cysteine synthase and ATPase functions, and structures for sequences specific to *T. cruzi* and absent in *H. sapiens* associated with 2,4-dienoyl-CoA reductase, and leishmanolysin activities. The implications of our structures refined by atomistic molecular dynamics (monomer or dimer states) in their *in vitro* environments (aqueous solution or membrane bilayers) are discussed for drug development and suggest that all protein targets, except cysteine synthase, merit further investigation. All these findings are summarized in Chapter V, which has been subject of an article: Lima CR, Carels N, Guimaraes AC, Tufféry P, Derreumaux P. *In silico structural characterization of protein targets for drug development against Trypanosoma cruzi.* J Mol Model 2016 Oct; 22(10): 244.

In the last Chapter, I present the main final conclusions with some on-going projects. Finally the Annex reports an article to be submitted.
Aims of the Thesis


Materials and Methods

CHAPITRE IV
Finding the lowest free energy (native) conformation of proteins and their folding mechanisms from computer simulations are two theoretical challenges in modern structural biology. Protein systems are very challenging due to their complex organizations (from a single chain to multiple chains), the wide range of relevant time scales ranging from microseconds to days, and the fact the majority of important processes depend on weak, non-covalent interactions that are only marginally stable at room temperature. All these characteristics require continuous progress in many aspects of molecular simulations [1-3].

Under in vitro and in vivo conditions, proteins have well-defined 3D structures, though this paradigm is challenged by *intrinsically disordered proteins*, which are expected to represent 30% of eukaryotic genomes encoded proteins with wholly or partially disordered structures, where folding is intimately coupled to binding to their biological targets [4], and *amyloid proteins* which self-assemble into amyloid fibrils with a common cross-beta structure, despite no resemblance between their amino acid sequences [5]. On the experimental side, in spite of constant efforts, structure determination is difficult for large-size systems by NMR and X-ray diffraction.

Standard atomistic molecular dynamics (MD) simulations and generalized ensemble simulations such as replica-exchange MD (REMD) or simulated tempering (ST) in explicit aqueous solution are able to explore protein dynamics and thermodynamics, respectively [3]. Independently of the accuracy of all-atom force fields (CHARMM, AMBER, and GROMOS variants) and water force fields (TIP3P, TIP4P, SPC/E), generating protein dynamics on timescales beyond 5 microseconds is still a tour de force by using standard computer resources. Also determining the equilibrium configuration ensemble and converged thermodynamic properties such as the variation of the heat capacity or the entropy as a function of temperature is even more problematic, irrespective of the enhanced sampling method used to overcome the multiple-minima problem (e.g., temperature or Hamiltonian in REMD, ST or metadynamics variants) [6]. So very often, the calculated free energy surface in explicit solvent is questioned in terms of convergence.

Very recently, it has been possible to build a computer with processors specifically designed to perform standard molecular dynamics. This is the case of the Anton computer conceived by D.E Shaw in New York at Columbia University. In contrast to most
Materials and Methods

computers, Anton can only perform one task, a molecular dynamics simulation, but with a very high efficacy, almost one hundred times faster than the best supercomputer. With the specially MD-designed Anton computer, it has been possible to break the millisecond barrier and gain insights into the mechanisms, thermodynamics and kinetics of the folding of diverse monomeric proteins with 10–80 amino acids [7,8]. Anton has also been useful in the development of allosteric inhibitors that target previously unknown binding sites [9]. However, the dynamic processes of life at the molecular level require knowledge of the structure, dynamics and thermodynamics of biomolecules in a cellular and crowded environment [2].

While both Graphic Processor Units (GPU) and distributed simulations performed on a network of worldwide computers (such as Folding@home) can reduce the computer time, and the emergence of quantum computers will certainly change the field of computer simulations in 10-20 years, there are two currently used bioinformatics methods to predict protein structures.

The first generic method is de novo in character and uses the concept of assembling 3D fragments of various lengths and structures. This is the basic ingredient of the programs Rosetta [10,11], Pep-Fold [12] and Quark [13], and here I briefly describe Rosetta and Pep-Fold. The state-of-the art program, Rosetta, uses a low-resolution five-bead representation that uses the backbone heavy atoms and a centroid located at the side-chain center of mass, followed by an all-atom representation. Rosetta uses a Monte Carlo fragment assembly of length 9 and then three residues to sample a series of scoring functions with increasing levels of chemical detail in order to identify native states. Overall, the low-resolution energy function includes van der Waals hard sphere repulsion, environment, pair, side-chain packing density, contact order, secondary structure packing [helix–helix pairing, helix–strand pairing, strand–strand pairing, strand pair distance/register and strand arrangement into sheets], radius of gyration energetic contributions, and Ramachandran torsion angle filters. Additional hydrogen bonding (short- and long-range backbone–backbone hydrogen bond) energy terms are used during full-atom refinement. The Rosetta procedure enables structure prediction with atomic level accuracy of small proteins, but for larger proteins, the native state is virtually never explored. Also, it has been shown that Rosetta’s coarse-grained potentials may actually impede accurate structure prediction at full-atom resolution in that the lowest-energy structures are not necessarily the most native-like [14].
Materials and Methods

On the other hand, Pep-fold relies on a hidden Markov model derived structural alphabet (SA) of 27 letters to describe proteins as a series of overlapping fragments of four amino acids. The SA letters can be assimilated to a generalized secondary structure, extending the number of states from 4 (α-helix, coil, turn or bend, and β-strand) to 27, but not all transitions are possible between two consecutive letters, reducing therefore the size of the conformation search. Pep-fold consists of three steps. First, Pep-fold predicts a limited set of SA letters at each position from the sequence, and then performs a progressive assembly of the prototype fragments associated with each selected SA letter using a greedy algorithm driven by the coarse-grained OPEP force field. Note, that OPEP model, designed prior to Rosetta, uses a backbone representation consisting of Cα, N, C, O and H atoms and one bead for each side-chain. As Pep-fold uses a rigid assembly, it is necessary to smooth the OPEP side chain–side chain potential. The third step in Pep-fold refines the CG models by Monte-Carlo before generating atomistic models with full atom representation of the side-chains and performing a clustering of all models returned by the simulations [12].

By utilizing a benchmark of 56 peptides with 25-52 amino acids adopting various 3D topologies in aqueous solution at pH7, free of any cofactors or ions, it has been shown that Pep-fold2 and Rosetta lead to similar results for peptides containing β and α/β structures, but Pep-fold2 has a much higher prediction rate for peptides containing α-helical structures [12] as shown in Figure 25.

Figure 25. PEP-FOLD2 and Rosetta models. (green) Experimental conformation (pale green correspond to residues not in the rigid core). (cyan) Best model generated. (magenta) Best model returned in the five first ranks. (top) Models generated for 2l0g and 1usd. (middle, bottom) Models generated by PEP-FOLD2 and Rosetta for 1w4e and 1n9d (D’apres Shen et al. 2014).
Materials and Methods

Recently, a new version of Pep-Fold, version 3.0 was designed, and unlike previous versions of Pep-Fold, we now use in the second step the Forward Backtrack algorithm or a Taboo Sampling algorithm to generate sub-optimal series of states or trajectories. Each series of states then leads to the generation of one conformation using a rigid assembly procedure of prototype fragments. As previously, once the peptide is complete using OPEP, it is refined using 30 000 Monte-Carlo steps, prior to full-atom structure generation. This much faster procedure, about 50-100 times faster than Rosetta, while preserving the accuracy of the obtained conformations, opens new perspectives in exploring the possible conformations of 10,000 peptides per day [15,16].

The analysis of CASP11 results, CASP standing for Critical Assessment of protein Structure Prediction, shows that there is a modest improvement, over the years, of the quality of the predicted conformations for sequences, having low sequence identity (< 30%) with protein sequences with known 3D structures. Similarly, CASP11 has clearly shown that it is preferable to use homology modelling, rather than de novo or ab initio methods, for long sequences, independently of the sequence identity. We now present the basic foundations of homology modelling, which is by far the most widely used computational approach to predict the 3D structures of proteins. Note that almost all protein structure prediction servers rely chiefly on homology modelling, as seen in the community-wide blind benchmark CASP [17].

Homology modelling consists of four steps (see Figure 26): (i) Finding homologous template proteins of known structure, (ii) Selecting the best template or set of templates, (iii) Optimizing the multiple sequence alignment (MSA) between the query and template protein sequences, and (iv) Building the homology model for the query sequence that resembles as closely as possible the structures of the templates, accommodating for deletions and insertions of query residues with respect to the template structures.
Materials and Methods

Figure 26: A flow chart of the steps involved in comparative protein structure modeling

During the last 15 years, much progress has been made regarding the sequence-based steps 1 to 3. This is mainly owed to the development of more sensitive and accurate methods for sequence searching and alignment that compare sequence profiles or profile hidden Markov models (HMMs) with each other. In contrast, improvements of the last step have been marginal. In my thesis, I have searched for template 3D structures using the HHsearch method hosted on Mobyle against the PDB and SCOP databases [18]. Then 3D models were generated by HHalign-Kbest procedure [19] that performs HHM alignments based on primary sequences and 3D structures. It has to be noted that if HHalign-Kbest identified a hit with a sequence identity >35% (or < 35%), 20 (or 50) models were automatically created. Outputs consist of the five best models ranked according their Z-scores awarded by Qmean4 [20] and only the best model was used for homology modeling.

A first important aspect in homology or template-based modeling is to evaluate the quality of the generated 3D models. In my thesis, the quality of the models was evaluated by using the meta-server PSVs 1.5 which analyzes the Ramachandran plot and provides
Materials and Methods
Molprobity and Procheck scores [21]. Following CASP10 recommendations, models were also investigated by ModFOLD4 [22], which returns the estimation of both the global and local (per-residue) quality of 3D protein models.

A second important aspect in homology modeling is to refine the starting structure by all-atom MD simulations. In my thesis, simulations were performed with Charmm22* force field [8] at pH 7, i.e., with deprotonated Glu and Asp and protonated Arg and Lys for all proteins, and neutral His with a protonated N atom for all proteins except leshmanolysin. This all-atom force field has been able to fold properly many soluble proteins [8]. The soluble proteins were centered in a cubic box of TIP3P water molecules with a box extended 1.5 nm outside the protein on all sides, and the appropriate numbers of Na+ and Cl− ions were added to ensure neutral systems. The leshmanolysin protein simulation includes the Zn2+ ion covalently bonded to the three histidines of the protein by application of a harmonic restraint on the bond lengths connecting Zn2+ to His. The ATPase transmembrane protein was immerged in a 1-palmitoyl-2-oleoylphosphatidylethanolamine (POPE) bilayer and then centered in a cubic box of TIP3P molecules. Each system was subjected to energy minimization with steepest descent and TNPACK, followed by a 1 ns MD simulation in NPT ensemble [39], and a 1 ns NVT MD simulation with a velocity-rescaling thermostat found to sample the canonical ensemble.

In my thesis, the GROMACS program (version 5.03) was used to perform the MD simulations with periodic boundary conditions. The bond lengths with hydrogen atoms were fixed with the LINCS algorithm and the equations of motion were integrated with a time step of 2 fs. The electrostatic interactions were calculated using the particle mesh Ewald method and a cutoff of 1.1 nm. A cutoff of 1.2 nm was used for the Van der Waals interactions. The nonbonded pair lists were updated every 10 fs. In my thesis, the analysis was based on MD simulation of 100 ns at 300 K. The MD-generated structures were evaluated with respect to the minimized structure using the Cα root-mean square deviation (RMSD), and clustered using a RMSD cutoff of 3 Å.

While evaluating the RMSD over the full sequence is important, it is also crucial (i) to evaluate the stability of the residues in the binding site in bound state, (ii) to see reasonable contacts between those designed protein receptors and the natural ligands if available from the PDB templates, and (iii) to determine and compare the hydrophobicity analysis and electrostatic potential of active sites of the static generated model and all MD-
Materials and Methods

generated structures with the homologs. In my thesis, hydrophobicity analysis was performed using the Kyte and Doolittle scale [23], as implemented in Chimera [24], and electrostatics were calculated using the PDB2PQR on-line facility, which performs an analysis of Poisson-Boltzmann electrostatics calculation [25], i.e. the most accurate method among several alternatives.

A final aspect in homology modeling and comparing two 3D models between two species is to determine (and this rarely done in CASP) whether the two models lead to identical normal mode analysis in spite of the insertions and deletions. In my thesis, I employed the Elnemo [26] and WEBnm@ [27] on-line services.
Materials and Methods


13. Xu D, Zhang Y Ab Initio Protein Structure Assembly Using Continuous Structure
Materials and Methods


Materials and Methods


CHAPlTRE V
In silico structural characterization of protein targets for drug development against *Trypanosoma cruzi*

Carlyle Ribeiro Lima¹², Nicolas Carels³, Ana Carolina Ramos Guimaraes⁴, Pierre Tufféry¹, Philippe Derreumaux²

¹Molécules Thérapeutiques in silico, Université Paris Diderot, INSERM UMR-S 973, Sorbonne Paris Cité, 75013, Paris, France
²Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico – Chimique, Université Paris Diderot, Sorbonne Paris Cité, 75005, Paris, France
³Laboratório de Modelagem de Sistemas Biológicos, National Institute for Science and Technology on Innovation in Neglected Diseases (INCT-IDN), Centro de Desenvolvimento Tecnológico em Saúde (CDTS), Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, RJ, Brazil
⁴Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz (IOC), FIOCRUZ, Rio de Janeiro, Brazil

Corresponding authors: pierre.tuffery@univ-paris-diderot.fr, philippe.derreumaux@ibpc.fr

**ABSTRACT**

*Trypanosoma cruzi* is the protozoan pathogen responsible for Chagas disease, which is a major public health problem in tropical and subtropical regions of developing countries and particularly in Brazil. Despite many studies, there is no efficient treatment against Chagas disease, and the search for new therapeutic targets specific to *T. cruzi* is critical for drug development. Here, we have revisited 41 protein sequences proposed by the analogous enzyme pipeline, and found that it is possible to provide structures for *T. cruzi* sequences with clear homologs or analogs in *H. sapiens* and likely associated with trypanothione reductase, cysteine synthase and ATPase functions, and structures for sequences specific to *T. cruzi* and absent in *H. sapiens* associated with 2,4-dienoyl-CoA reductase, and leishmanolysin activities. The implications of our structures refined by atomistic molecular dynamics (monomer or dimer states) in their *in vitro* environments (aqueous solution or membrane bilayers) are discussed for drug development and suggest that all protein targets, except cysteine synthase, merit further investigation.

**Keywords:** T. Cruzi, protein targets; structures, molecular dynamics, drugs.
1. Introduction

According to the World Health Organization, 21 Latin American countries are endemic for Chagas disease, which is caused by the parasite Trypanosoma Cruzi (T. cruzi), and about 6 million to 7 million people are infected with T. cruzi [2]. Chagas disease has two phases: the initial acute phase lasts for about 2 months with mild or little symptoms, followed by the chronic phase, in which 30% (10%) of patients suffer from cardiac (digestive) disorders before death or heart failure [1].

Chagas disease can be cured with benznidazole and nifurtimox if treatment is initiated soon after infection [3]. Both drugs are activated by a NADH-dependent, mitochondrially localized, bacterial-like, type I nitroreductase (NTR) and resistance to these compounds is associated with loss of NTR [4]. It is well established that both drugs are far from being ideal because of their controversial efficacies in chronic phase, severe host toxicity and long treatment durations [5].

The identification of new protein targets for drug development against Chagas disease is therefore needed and strategies to identify enzymatic functions specific to the particular T. cruzi pathogen are highly desirable. Indeed, the enzyme forms that are absent in humans can be used as targets for the development of new drugs that would have a low chance of side effects [6-8].

Recently, several proteins have been proposed as new biotargets against Chagas disease and 3D models were built based on homology modeling followed by molecular dynamics simulations. These T. cruzi biotargets include NADH-dependent fumarate reductase [9], ribosomal P0 antigenic protein [10], silent-information regulator 2 proteins or sirtuins [11], and a putrescine–cadaverine permease [12], among others. While these studies provide useful information on specific targets, a large-scale search of promising proteins is desirable.

In this context, the analogous enzyme pipeline (AnEnPi), which compares genomic datasets for analogous enzymes by clustering the primary structures of enzymes with the same described activity, is a tool able to identify enzymatic activities that may be potentially useful for the design of new therapeutic targets. AnEnPi searches for enzymes specific to a parasite and enzymes that are analogous for the same enzymatic function, but structurally distinct in the parasite and its human host [13]. Using AnEnPi with a Blastp similarity raw score of 120 as cut-off, Capriles et al. identified 41 protein sequences classified as analogous or specific for T. cruzi compared to H. sapiens on the basis of
Article

similarity detection by the local alignment program BLAST, and then used automatic 3D structure modeling to infer differences between the host and parasite protein structures [14].

In this study, we revisit these 41 protein candidates both in terms of specificity and 3D modeling by using a more complex strategy and more accurate techniques. Indeed, it is known that profile-profile alignments are more sensitive than the local alignment program PSIBLAST for similarity [15,16], and automatic 3D modeling methods lead to significant structural errors for sequence identities < 80% [17]. The workflow for T. cruzi protein modeling we developed consists in the following steps: sequence clustering, prediction of cell localization and enzyme classification (EC) number and analysis of metabolic pathways, homology search in H. sapiens and other organisms, 3D template identification, model generation and refinement by molecular dynamics (MD) simulations refinement.

Overall, we find that it is possible to cluster the 41 protein sequences identified as analogous or specific for T. cruzi versus H. sapiens into 7 groups, and based on the current Protein Data Bank (PDB) structures, it is possible to provide structures for 33 T. cruzi sequences associated with five very probable enzymatic activities: ATPase (EC: 3.6.3.6), trypanothione reductase (1.8.1.12), 2,4-dienoyl-CoA reductase (1.3.1.34), cysteine synthase (2.5.1.47), and leishmanolysin (3.4.24.36). The implications of our results for drug development are discussed.

2. Materials and methods

Our pipeline for the 41 protein sequences is shown in Fig. 1.

**Sequence clustering.** We used the sequence search of the PFAM server to identify the PFAM domains for of each of the 41 sequences [18]. Sequences having the same PFAM domains were then grouped into clusters.

**Cell localization prediction.** We used both the EUK-mPloc 2.0 and CELLO v.2.5 servers [19,20]. Note that the EUK-mPloc 2.0 learning dataset includes 8,897 protein sequences (7,766 different proteins), classified into 22 eukaryotic subcellular locations

**EC number prediction and metabolic pathway identification.** Pathway graphs can be thought as three information layers: (i) the lowest layer is associated with the metabolites that act as substrates (input) and products (output) of enzymatic reactions, (ii)
the intermediate layer consists of the enzymes that perform the enzymatic reaction with which the biological engineer intended to interfere through drug inhibition, and (iii) the highest organization level is given by the topology of a metabolic path in which enzymes are inserted. An enzyme within a non-redundant path is a better target than an enzyme in a redundant one since its inactivation will inevitably resulted in the elimination of the corresponding product. If that path is a key path for the whole parasite metabolism, the target inhibition is expected to have deleterious consequences on the parasite biology. Considering these three organization levels [21], we used KEGG [22,23] to map the EC number of an enzyme target and MetaCyc [24] to retrieve information about the metabolic reactions of each enzymatic reaction. We also checked that our predictions are consistent with AnEnPi server [13].

**Homology identification.** To identify homologs in *H. sapiens* and other organisms we have used HHpred [25]. A homolog was identified if the sequence identity is $>30\%$ for an amino acid coverage of more than 90%.

**3D modeling of *T. cruzi* targets.** We searched for template 3D structures using the HHsearch method hosted on Mobyle against the PDB and SCOP databases [26]. Then 3D models were generated by HHalign-Kbest procedure [27] that performs HHM alignments based on primary sequences and 3D structures. It has to be noted that if HHalign-Kbest identified a hit with a sequence identity $>35\%$ (or $<35\%$), 20 (or 50) models were automatically created. Outputs consist of the five best models ranked according their Z-scores awarded by Qmean4 [28] and only the best model was used for homology modeling.

When the template did not cover the full sequence, we examined the putative biological information of the protein (aqueous solution, anchored or inserted into the membrane), and then used the template-based C8-Scorpion [29], the Porter 4.0 [30] and PrDOS [31] servers to determine whether the missing residues are predicted to have secondary structure content or disordered.

**Model refinement by molecular dynamics.** Simulations were performed with Charmm22* force field [32] at pH 7, i.e., with deprotonated Glu and Asp and protonated Arg and Lys for all proteins, and neutral His with a protonated N\[\square\] atom for all proteins except leishmanolysin. This all-atom force field has been able to fold properly many soluble proteins [33-35]. The soluble proteins were centered in a cubic box of TIP3P water
molecules [36] with a box extended 1.5 nm outside the protein on all sides, and the appropriate numbers of \(\text{Na}^+\) and \(\text{Cl}^-\) ions were added to ensure neutral systems. The leshmanolysin protein simulation includes the \(\text{Zn}^{2+}\) ion covalently bonded to the three histidines of the protein by application of a harmonic restraint on the bond lengths connecting \(\text{Zn}^{2+}\) to \(\text{His}\). The ATPase transmembrane protein was immersed in a 1-palmitoyl-2-oleoylphosphatidylethanolamine (POPE) bilayer and then centered in a cubic box of TIP3P molecules [37]. Each system was subjected to energy minimization with steepest descent and TNPACK [38], followed by a 1 ns MD simulation in NPT ensemble [39], and a 1 ns NVT MD simulation with a velocity-rescaling thermostat found to sample the canonical ensemble [40].

The GROMACS program (version 5.03) was used to perform the MD simulations with periodic boundary conditions [41]. The bond lengths with hydrogen atoms were fixed with the LINCS algorithm and the equations of motion were integrated with a time step of 2 fs [42]. The electrostatic interactions were calculated using the particle mesh Ewald method and a cutoff of 1.1 nm [43]. A cutoff of 1.2 nm was used for the Van der Waals interactions. The nonbonded pair lists were updated every 10 fs. In what follows, the analysis is based on MD simulation of 100 ns at 300 K. The MD-generated structures were evaluated with respect to the minimized structure using the \(\text{C}^\alpha\) root-mean square deviation (RMSD), and clustered using a RMSD cutoff of 3 Å.

**Model evaluation.** The quality of the models was evaluated by using the meta-server PSVs 1.5 which analyzes the Ramachandran plot and provides Molprobity and Procheck scores [44]. Following CASP10 recommendations [45], models were also investigated by ModFOLD4 [46], which returns the estimation of both the global and local (per-residue) quality of 3D protein models.

Hydrophobicity analysis of active sites was performed using the Kyte and Doolittle scale [47], as implemented in Chimera [48], and electrostatics were calculated using the PDB2PQR on-line facility, which performs an analysis of Poisson-Boltzmann electrostatics calculation [49].

For two protein targets, we also examined the vibrational frequency modes of *T. cruzi* and *H. sapiens* models by using the Elnemo [50] and WEBnm@ [51] on-line services.
3. Results and Discussion

Table S1 in Supplementary Materials presents the 41 targets considered in this study, with their identification numbers in *T. cruzi* genome, amino acid lengths, family descriptions, PFAM domains and predicted EC numbers. Our predicted EC numbers are consistent with the original genome function annotation list from GeneDB [52]. Using our methodology, we are able to cluster these 41 sequences identified by Capriles et al. as analogous or specific for *T. cruzi* versus *H. sapiens* into 7 groups based on PFAM analysis and enzymatic activity. The group 4 of 8 sequences related to cruzipain (EC: 3.4.22.51) and the group 5 related to triacylglycerol lipase (EC: 3.1.1.3) with 1 sequence are eliminated after PFAM protein domain analysis because they include large domains (> 250 amino acids) of unknown functions (e.g. DUF3586 in cruzipain) that lack any sequence homology with available PDB structures and cannot be modeled by *de novo in silico* methods [53-55] at the moment. The remaining five groups are related to ATPase (EC: 3.6.3.6, 2 sequences), Trypanothione disulfide reductase (EC: 1.8.1.12, 2 sequences), 2,4-dienoyl-CoA reductase (EC: 1.3.1.34, 2 sequences), leishmanolysin (EC: 3.4.24.36, 23 sequences), and cysteine synthase (EC: 2.5.1.47, 2 sequences). In the first three groups, as one sequence is a fragment of the other, we analyzed the sequence of larger size. Both cysteine synthase proteins have 332 residues and share 98% sequence identity, and one sequence was modeled. The 23 sequences in the leishmanolysin group have between 516 and 621 amino acids, and as explained below, we modeled the structure of one sequence.

Table 1 summarizes the main results of our procedure for the five sequences considered for 3D structure and refinement. For each target, we give its predicted cell localization, whether we found significant homology hit with *H. sapiens* and distant organisms in terms of sequences and PDB structures, and finally the amino acid alignment between the *T. cruzi* and homologous structures. Among these five sequences, only the trypanothione disulfide reductase and the cysteine synthase sequences have clear homologs in *H. sapiens*. We now describe the details of our analysis for each protein and its implication for drug design.

**ATPase (predicted EC: 3.6.3.6).** The *T. cruzi* ATPase sequence of 898 amino acids has a clear homolog (36% sequence identity) with an H⁺-transporting ATPase in *Arabidopsis Thaliana* (PDB: 3B8C [56], region 7-885, EC 3.6.3.6). The *T. cruzi* sequence has 26% sequence identity with an ATPase in *Oryctolagus cuniculus* (PDB: 3AR4 [57], region 1-994, EC: 3.6.3.8), and 23% with a Na⁺/K⁺ transporting ATPase in *Mus musculus*
(region 9-851, EC: 3.6.3.9). There is no clear homolog in *H. sapiens*, as there is at most 24% sequence identity with a calcium transporting ATPase (region 7-872, EC: 3.6.3.8).

Using the X-ray structure from *A. thaliana* (PDB: 3B8C) as a template for *T. cruzi* and *H. sapiens* sequences, there is a Cα RMSD of 1.2Å between the *T. cruzi* model and the template, and 1.0Å between the *T. cruzi* and *H. Sapiens* structures. Figure S1 shows the superposition of the *T. cruzi* model on the 3B8C template. However, among five residues important for ACP (phosphomethylphosphonic acid adenylate ester) binding in 3B8C, three residues are conserved (F385 (453), K408 (479) and R441 (522) in *T. cruzi* (*H. sapiens*) amino acid numberings), but L443 in *T. cruzi* is changed to G517 in *H. sapiens*, i.e. changing the flexibility of the backbone, and more importantly, the K423 residue in *H. sapiens* is mutated to D359 in *T. cruzi*, i.e. changing the electrostatic feature of the active site. This makes the choice of this template questionable.

Because the X-ray structure from *A. thaliana* has a very low resolution (3.6 Å), the X-ray structure from *Oryctolagus cuniculus* (PDB: 3AR4) with a resolution of 2.15Å is definitely a better structural template and the sequence alignment shown in Figure S2 was used for the homology modeling and MD refinement. Fig. 2A shows the superposition of the *T. cruzi* structure on the 3AR4 template with a Cα RMSD of 0.2Å (free of any minimization) with the location of the six transmembrane α-helices (residues 50-99, 219-284, 614-644, 664-725, 749-799 and 823-854 in *T. Cruzi*). Stability of the *T. cruzi* structure was assessed by 100 ns MD simulation at 300 K, which led to a mean RMSD of 5.3Å from the minimized structure (Fig. 2D). Figures 2E shows the superposition of the most populated cluster representing 85% of the conformational MD ensemble on the minimized structure. The minimized and all MD generated structures have good but lower quality than the structure arising from comparative modeling using Ramachandran and ModFOLD4 metrics (Fig. S3).

The *T. cruzi* sequence is predicted to be a P-type H⁺-exporting ATPase (EC: 3.6.3.6) found in the oxidative phosphorylation pathway and allowing proton transportation across plasma membrane to generate the electrochemical potential gradient of cells. P-type ATPases catalyze the selective active transport of ions like H⁺, Na⁺, K⁺, Ca²⁺, Zn²⁺, and Cu²⁺ across diverse biological membrane systems, and several molecules have been shown to inhibit ATPase activity to different degrees [58]. For instance, artemisin was found to inhibit growth of cultured *T. cruzi* at concentrations in the low micromolar range and inhibit Ca²⁺-dependent ATPase activity in *T. cruzi* membrane [59]. Also miltefosin was
found to inhibit the Na\(^+\)-ATPase. This compound was however also found to inhibit the protein kinase C present in the plasma membrane of \textit{T. cruzi}[60].

The alignment of the \textit{T. cruzi} and \textit{H. Sapiens} structures (RMSD of 1.2Å) using \textit{O. cuniculus} template suggests that the machinery for the dynamic and rotational motion of the cytoplasmic-compartment sector with respect to the membrane-sector, driven by the H\(^+\) electrochemical potential gradient, is conserved. This is supported by normal mode analysis of the two systems, free of ATP ligand, since the lowest frequency modes are almost identical (data not shown). Fig. 2B zooms the active sites in the template, \textit{T. cruzi} and \textit{H. sapiens} and shows their hydrophobic and electrostatic surfaces. As seen, while the hydrophobic surface is slightly changed (bottom panel), the electrostatics potential (middle panel) is completely changed from \textit{T. cruzi} to \textit{H. sapiens}, being more negative in \textit{H. sapiens} than in \textit{T. cruzi}. Fig. 2C shows that four residues of the active site crucial for ATP binding are strictly conserved among the template, \textit{T. cruzi} and \textit{H. sapiens}, i.e., F385, K408, R441 and L443 in \textit{T. cruzi} amino acid numbering. As seen in Fig. 2B (bottom panel), the mutated residues between \textit{T. cruzi} and \textit{(H. sapiens)} are D359 (K423), L361 (T425), and T392 (M460). Interestingly, these three mutations observed in \textit{T. cruzi} are strictly conserved in \textit{T. rangeli}, \textit{T. vivax}, \textit{T. brucei}, \textit{T. congolense}, \textit{B. saltans}, \textit{L. major}, \textit{T. grayi} and \textit{G. theta}, indicating a strong evolutionary pressure of these positions in related species.

Whether these three mutations in or near the active site that change the electrostatic potential can govern the selectivity of the ATP-driven pumps for a drug in \textit{T. cruzi} with respect to \textit{H. sapiens}, and make the ATPase a suitable candidate, as suggested by the chemogenomics resource for neglected tropical diseases (TDR) [61] and cell surface proteome analysis of human-hosted \textit{T. cruzi} life stages [62], remains to be explored.

**Trypanothione Reductase (predicted EC: 1.8.1.12).** The 492 amino acid trypanothione reductase (TryR) in \textit{T. cruzi} has a clear homolog in \textit{Trypanosoma brucei} (PDB: 2WPF [63], EC: 1.8.1.12) with a sequence identity of 82% over the amino acid region 1-488. The \textit{T. cruzi} sequence has also 42% and 39% sequence identities with two glutathione disulfide reductase with EC: 1.8.1.7 in \textit{Pseudomonas aerugina} and \textit{Nostoc punctiform}, and 36% sequence identity with a glutathione reductase (EC: 1.8.1.7) in \textit{H. sapiens}. As a result, the X-ray structure (PDB: 2WPF), consisting of a homo-dimer with 2 x 488 amino acids, was selected as structural template and the sequence alignment shown in Figure S4 was used for the homology modeling of the \textit{T. cruzi} and \textit{H. sapiens}
sequences. The superposition of the 3D models created by HHalign-Kbest for *H. sapiens* and *T. cruzi* on the template leads to small RMSD deviations of 1 and 0.6 Å, respectively (Fig. 3A-B). Analysis of the 100 ns MD simulation of the *T. cruzi* model at 300 K shows that the system is very stable (Fig. 3D), with one unique cluster and a mean RMSD of 2.5 Å from the full homo-dimer minimized structure (Fig. 3E). As for ATPase models, the MD generated structures of trypanothione reductase have lower quality than the structure obtained by comparative modeling as evaluated by the Ramachandran and ModFOLD4 metrics (Fig. S3). Under the harmonic approximation, the *T. cruzi* and *H. sapiens* structures show identical atomic fluctuations and deformation energies projected onto the calculated normal modes along the amino acid sequence (data not shown).

Trypanothione reductase is an essential enzyme of the unique trypanothione-based thiol metabolism of Trypanosomatidae. Trypanothione is an unusual form of glutathione containing two molecules of glutathione joined by a spermidine (polyamine) linker. TryR is a flavoenzyme protein that catalyzes the reaction trypanothione + NADP⁺ ↔ trypanothione disulfide + NADPH + H⁺ [64]. Trypanoso-momatids lack both glutathione reductase and thioreductase, and as a result, TryR is the only connection between NADPH- and thiol-based redox systems, the latter substituting for many antioxidant functions [65].

The active sites in *T. cruzi* and *T. brucei* have both hydrophilic and hydrophobic characters, while the active site in *H. sapiens* is highly hydrophobic (Fig. 3C bottom). Further analysis of the active sites shows that all residues binding the WPF ligand (3,4-dihydroquinazoline inhibitor) are conserved between *T. cruzi* and *T. brucei*, while only a single cysteine at positions 53 and 42 is conserved between *T. cruzi* and *H. sapiens* (Fig. 3C top), changing therefore drastically the electrostatic potential surface (Fig. 3C middle. Clearly, the low amino acid conservation of the catalytic site between trypanosomatidae and humans, and the importance of this enzyme for protozoan parasites, make this protein target very attractive for drug development.

It has to be emphasized that previous studies have reported *T. cruzi* trypanothione reductase inhibitors such as the antimicrobial chlorhexidine and a piperidine derivative [66]. Recently, Lavorato et al. described the antitrypanosomal activity and cytotoxicity profile of 20 novel 1,3-bis(aryloxy)propan-2-amine derivatives as new candidates for further development as potential anti-trypanosomal agents [67]. Also 82 novel TryR inhibitors down to the nM range were identified by using a combined *in vitro* and *in silico* screening approach [68].
2,4-dienoyl-CoA reductase (predicted EC: 1.3.1.34). The 717 amino-acid *T. cruzi* protein shares 31% sequence identity with the 2,4-dienoyl-CoA reductase (DECR1, EC: 1.3.1.34) from *E. coli* (region 1-671) with a structure determined by X-ray (PDB: 1PS9) at a 2.2Å resolution [69]. The cellular localization and EC number in *T. cruzi* are predicted to be identical to those in *E. coli*. The DECR1 enzyme participates in the beta-oxidation and metabolism of polyunsaturated fatty enoyl-CoA esters. DECR1 catalyzes the reaction \[
\text{trans-2,3-Dehydroacyl-CoA} + \text{NADP}^+ \leftrightarrow \text{trans,trans-2,3,4,5-Tetradehydro acyl-CoA} + \text{NADPH} + \text{H}^+.
\]

The *T. cruzi* sequence has also 34% sequence identity with a protein in *Pseudomonas aeruginosa* with the same EC number (1.3.1.34). Searching in the proteome of *H. sapiens*, the *T. cruzi* sequence shares only 19% sequence identity with two proteins differing by EC numbers (4.2.1.22 and 4.1.1.17) and covering only the amino acid region 391-717.

Using the sequence alignment shown in Fig. S5 for homology modeling, the superposition of the *T. Cruzi* model (sequence 4-717) on the 1PS9 structure (sequence 1-671) has a RMSD of 0.3Å prior to energy minimization (Fig. 4A). Analysis of the 100 ns MD simulation at 300 K shows that the *T. cruzi* model fluctuates about 5Å from its minimized structure (Fig. 4C), and the large RMSD deviation comes from the long loop regions that need to be introduced from residues 666 to 717 (Fig. S5). Analysis of the MD-generated structures leads to 2 clusters, with the first cluster representing 90% of the ensemble (Fig. 4D) and high quality as assessed by ModFOLD4 values (Fig. S3).

There are currently no chemical compounds associated to the *T. cruzi* gene, as reported in the TDR database. Most catalytic residues that allow FMN (Flavin mononucleotide) substrate binding are conserved between the *T. cruzi* and *E. coli* sequences (G57, Q100, R214, R288, A330 and R331 in *E. coli* vs. G68, Q111, R230, R310, A332 and R333 in *T. cruzi*) and lead to a positive electrostatic field (Fig. 4B). The two residues S24 and H26 in *E. coli* are changed into P30 and Y32 in *T. cruzi* (Fig. 4B), but these two mutations are not sufficient to alter the polar features of the binding site, located in an inner enzyme region (Fig. 4C). Interestingly, humans have an enzyme for the same function, but with a completely 3D structure (homotetramer, each of 302 amino acids, PDB 1W6U), and a distinct catalytic center [70], making DECR1 very suitable for the design of drugs specific to *T. cruzi*. 

89
Cysteine synthase (predicted EC: 2.5.1.47). The *T. cruzi* sequence of 332 amino acids has a clear homolog (73% sequence identity) with a cysteine synthase of 324 amino acids from *Leishmania donovani* (region: 2-319, EC: 2.5.1.47, PDB: 3TBH at a 1.74 Å resolution [71]). The *T. cruzi* sequence also shares 55% sequence identity with a cysteine synthase in *Arabidopsis thaliana* (EC: 2.1.1.47) and 33% sequence identity with a human protein (region 6-332 with EC: 4.2.1.22, a cystathionine β-synthase). As a result, the homodimer structure from *Leishmania donovani* was selected as structural template and the sequence alignment shown in Figure S6 was used for the homology modeling of the *T. cruzi* protein. Figures 5A and 5B show the superposition of the predicted HHAlign-Kbest model on the *Leishmania donovani* structure with a RMSD of 0.2Å. Analysis of the 100 ns MD simulation at 300 K of the homodimer of *T. cruzi* shows an averaged RMSD deviation of 4.5Å with respect to the minimized structure (Fig. 5D), with the most populated cluster, representing 70% of the conformational ensemble (Fig. 5E) and with high quality as measured by the PSVS metrics (Fig. S3).

There is currently no chemical compound associated to the *T. cruzi* gene, as reported in the TDR database [61]. Cysteine synthase is a key constituent for the survival of trypanosomatids, and thus vital to the survival of *T. cruzi* in its vertebrate hosts. It constitutes one of the key pathways in the parasite defense against oxidative stress. Two different routes for cysteine biosynthesis have been described: reverse-transsulfuration (RTS) and *de novo* pathways. RTS has been demonstrated in fungi and mammals and includes the complete process leading to cysteine from methionine via the intermediary formation of cystathionine. These reactions are catalyzed by two enzymes: CβS (cystathionine β-synthase) and CGL (cystathionine γ-lyase). The *de novo* pathway, catalyzed by two steps starting with serine acetyltransferase to form O-acetylserine from L-serine and acetyl-coenzyme A, is found in plants, bacteria, and some protozoa, such as *Entamoeba histolytica*, *Leishmania major* and *Leishmania donovani*, but is absent in mammals [72].

*T. cruzi* sequence is a clear homolog of an O-acetylserine sulphydrylase from *L. donovani*. The active site with the Asp-Gly-Ser-Gly-Ile ligand is fully conserved in terms of amino acids and hydrophilic character from *T. cruzi* to *L. donovani*, with only a single amino acid change (T80 changed to S79) (Fig. 5C). As shown on Fig.S6, the residues involved in the active sites with cysteine synthase and cystathionine β-synthase activities are conserved in *T. cruzi*, *L. donovani* and *H. sapiens* and the differences in the physico-chemical properties of the residues flanking the active sites are minor (Fig. S7). Clearly,
the high conservation of the active site and the presence of this enzyme in humans do not make this protein a good target for drug design specific to *T. cruzi*. However, there have been several reports of inhibitor screening and development against this molecule from different organisms such as *Entamoeba histolytica* and *Mycobacterium tuberculosis*, using an off-catalytic site strategy, and it was suggested that among the open form, the intermediate state and the closed form of the enzyme, the intermediate state of the enzyme might be the ideal target for the design of very effective high-affinity inhibitors [73].

**Leishmanolysin (predicted EC: 3.4.24.36).** The superfamily of Leishmanolysin consists of 23 *T. cruzi* sequences with 516-621 amino acids and cross sequence identities varying between 37% and 96%. Multiple sequence alignment using ClustalW [74] shows that 22 sequences (except the smallest of 516 amino acids with a truncated N-terminus) have a common domain of 565 amino acids with > 30% sequence identity and, among these, there are 7 sequences of 567 amino acids and 7 sequences with > 570 amino acids displaying either longer N-terminus (2 sequences), longer C-terminus (4 sequences), or longer N- and C-termini (1 sequence). The N-terminus can be extended by up to 56 amino acids.

Taking the TC00.1047053511211.90 sequence of 567 amino acids as a representative of the 23 *T. cruzi* sequences, the region 77-527 shares 37% sequence identity with a Leishmanolysin from *Leishmania major* (region: 103-558, EC: 3.4.24.36, PDB: 1LML at a 1.8 Å resolution [75]). This sequence has at most a sequence identity of 25% with a protein from *H. sapiens* with EC= 4.2.1.22 (region 69-603), but the alignment shows a gap percentage of 16%. Based on this result, the X-ray structure 1LML was selected as structural template of the region 77-527, and it remained to determine whether the missing residues 1-76 and 528-567 could be modeled by other means. The N-terminus cannot be easily modeled since Leishmanolysin is a membrane-bound zinc proteinase, active in situ, but the exact anchorage of the N-terminus to the membrane is unknown. Using the best secondary structure prediction methods, i.e., C8-Scorpion and Porter 4.0, the region 528-567 is predicted to be free of any secondary structure. Similarly, the C-terminus region is predicted as disordered using PrDOS, preventing therefore its 3D modeling. Note that this intrinsic disorder of the C-terminus is shared by the other 21 sequences.
The sequence alignment shown in Figure S8 was used for the homology modeling, and the superposition of the *T. cruzi* structure on its template (with a RMSD of 0.4Å) is shown in given in Fig. 6A. Fig. 6B zooms the active site in the template and *T. cruzi* and shows the high hydrophilic character and negative electrostatic field of the active site. The three histidines essential for Zn$^{2+}$ binding are conserved in both species, as well as the residues Glu and Met. Stability of the *T. cruzi* model complexed by Zn$^{2+}$ was assessed by 100 ns MD simulation at 300 K. It is found that the system has an average RMSD of 5.5Å from the minimized structure (Fig. 6C), the most populated cluster representing 50% of the conformational MD (Fig. 6D) having very good quality using PSVS and ModFOLD4 metrics (Fig. S3).

Leishmanolysin is the predominant glycoprotein surface antigen of promastigotes of various species of *Leishmania*. The crystal structure has shown three domains, two of which had novel folds at that time, and the N-terminal domain has a similar structure to the catalytic modules of zinc proteases, revealing therefore that leishmanolysin is a member of the metzincin class of zinc proteinases [75]. The similarity of the active site structure and amino acid composition in *T. cruzi* model to previously well characterized metzincin class zinc proteinases suggests to test known zinc-metallopeptidase inhibitors, such as 1,10-phenanthroline or peptide-based inhibitors [76].

We would like to emphasize that for the five targets, we have evaluated the stability of the residues in the binding site in bound state by calculating the Cα RMS deviations of the active site residues in the minimized structure and the most populated (best) MD cluster with respect to their starting positions. All residues in the active site remain at their native positions although the ligands (ATP in ATPase, WPF in Trypanothione Reductase, FMN in NADPH and Asp-Gly-Ser-Gly-Ile in Cysteine Synthase) are absent in the simulations, with most residues deviating by less than 0.3 Å and at most by 1 Å. Also we have checked that we have reasonable contacts between the designed protein receptors and their natural ligands by analyzing the Cα and Cβ deviations of the active site residues in the *T. cruzi* minimized structure and most populated MD cluster with respect to the PDB structure used for homology without any minimization. Small values of 0.1-0.2 Å indicate that the contact maps for the designed *T. cruzi* minimized proteins binding to their corresponding ligands are very similar to the contacts maps of the homolog proteins. Largest values are expected and observed for the designed *T. cruzi* most populated MD cluster, but all remain below
Finally, we have verified that the dynamics does not change the hydrophobic/hydrophilic character and the electrostatic potential of the active sites, these properties being essentially conserved from the minimized structure to the most populated MD cluster.

4. Conclusion

The identification of specific \textit{T. cruzi} protein targets for drug development is desirable. In recent years, drug discovery for Chagas disease has evolved from limited compound testing in manual assays to sophisticated in vitro and \textit{in vivo} assays. Drug discovery must consider that \textit{T. cruzi} is a genetically diverse organism with phenotypically variations, and inhibitor susceptibility is a well-established phenotypic variation both \textit{in vitro} and \textit{in vivo} [77]. In the next five years, 200,000 people living with Chagas disease will die from heart disease and related complications, and the BENEFIT trial pursues a global initiative of diagnosis, treatment, and research [78].

In this study, by following a complex bioinformatics procedure, we have proposed structural models for five \textit{T. cruzi} enzymes with ATPase, trypanothione reductase, 2,4-dienoyl-CoA reductase, cysteine synthase, and leishmanolysin functions. Based on sequence and active-site similarity with several organisms and notably \textit{H. sapiens}, we show that these five protein targets, except cysteine synthase, are very attractive for drug development against \textit{T. cruzi}. This development based on computer simulations is not straightforward, however, because it is important to explore a much larger dynamic time-scale of the enzymes and enzymes/inhibitors using advanced simulation methods so as to evaluate the correct binding pose [79,80] and time-consuming free energy procedures in order to predict accurate binding affinity that goes beyond simple scores [81-83].

Conflict of interest

The authors declare no financial conflict of interest regarding the publication of this manuscript.

Acknowledgements

C.L. thanks CAPES/CsF and Science without Borders PhD grant program from Brazil, BEX Process 11917/13-0.

Appendix A.
Article

Supplementary data related to this article can be found in
References


Article


Table 1. Cluster representatives.
For each cluster (#), only one representative sequence is retained, and we report its T.
**Article**

cruzi DB identifier (the sequence can be accessed at TcruziDB.org), Pfam superfamily, size (L) in amino acids, predicted cellular localization, identified human / non human homolog sequences, the PDB template used for 3D modeling, and the limits of the aligned residues (coverage (Cov). – T cruzi / Template). For homolog sequences, we report the associated EC number, the sequence identity (%) and the limits of the homologous region in the T.cruzi sequence. For the template, we give the resolution, the organism, the Pfam superfamily, the EC number and the sequence identity (%).

<table>
<thead>
<tr>
<th>#</th>
<th>PFam superfamily</th>
<th>TcruziDB Id.</th>
<th>L</th>
<th>Cell. Localization</th>
<th>Human homologues</th>
<th>Non human Homologues</th>
<th>Template</th>
<th>Cov.*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(EC number, %identity)</td>
<td>(EC number, %identity, i coverage)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>ATPase</td>
<td>TC00.1047053508649.20</td>
<td>898</td>
<td>Plasma - membrane</td>
<td></td>
<td>*A. thaliana (3.6.3.3, 66% - 863)</td>
<td>3AR4 (2.1.5A) Oryctolagus cuniculus ATPase</td>
<td>4 – 855 /</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*Mus musculus (3.6.3.9, 23% - 861)</td>
<td>ATPase1</td>
<td>3.6.3.8</td>
</tr>
<tr>
<td>2</td>
<td>Trypanothione disulfide reductase</td>
<td>TC00.1047053503555.30</td>
<td>492</td>
<td>cytoplasmic</td>
<td>*P. aeruginosa (1.8.1.7, 42% - 76)</td>
<td>2WP2 (1.90A) Trypanosoma brucei</td>
<td>2 – 492 /</td>
<td>1 – 488</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*N. punctiforme (1.8.1.7, 39% - 2 – 469)</td>
<td>Trypanothione reductase</td>
<td>1.8.1.12</td>
</tr>
<tr>
<td>3</td>
<td>2,4-dienoyl-CoA reductase (NADPH)</td>
<td>TC00.1047053510303.210</td>
<td>717</td>
<td>cytoplasmic</td>
<td>*P. aeruginosa (1.3.1.34, 34% - 716)</td>
<td>IPS9 (2.20A) Escherichia coli 2,4-dienoyl-CoA reductase</td>
<td>1.3.1.34</td>
<td>31%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*A. fulgidus (1.3.1.34, 34% - 716)</td>
<td>Trypanothione reductase</td>
<td>1 – 717 /</td>
</tr>
<tr>
<td>6</td>
<td>Leishmanolysin</td>
<td>TC00.1047053511211.90</td>
<td>567</td>
<td>Plasma - membrane</td>
<td></td>
<td>C. elegans (3.4.24 - 276, 69 – 603)</td>
<td>ILML (1.86A) Leishmania major Leishmanolysin 3.4.24 - 376</td>
<td>77 – 527 /</td>
</tr>
<tr>
<td>7</td>
<td>Cysteine synthase</td>
<td>TC00.1047053507165.50</td>
<td>332</td>
<td>extracellular</td>
<td>cystathionine β-synthase (4.2.1.22, 33%)</td>
<td>A. thaliana (2.5.1.47, 55% - 1 – 320)</td>
<td>3TBH (1.74A) Leishmania donovani cysteine synthase 2.5.1.47</td>
<td>73%</td>
</tr>
</tbody>
</table>
Figures legends

**Figure 1:** Workflow used for protein characterization.

**Figure 2:** A) ATPase (EC 3.6.3.6). *T. cruzi* model (cyan), superimposed on *O. cuniculus* structure (PDB: 3AR4 - orange) in explicit membrane environment. B) (Top): Focus on the active site using the template (orange) with ATP (gray) of *T. cruzi* model (cyan) in right and *H. sapiens* model (green). (Middle): Electrostatic potential (Kb.T.ec$^{-1}$) – blue: negative, red: positive. (Bottom): Hydrophobicity – orange: hydrophobic, blue: hydrophilic. C) Varying amino acids of the active site in the three species. D) RMSD variation as the function of time (ns), relative to the initial structure. E) Superimposition of the initial structure (cyan) and most populated cluster (red) over 100 ns period molecular dynamics.

**Figure 3:** A) Dimer structure of the Trypanothione disulfide reductase (EC 1.8.1.12). *T. cruzi* model (cyan), superimposed on *T. brucei* structure (PDB: 2WPF - orange) and *H. sapiens* model (green). Gray depicts the second monomer. B) Focus on only Chain A of *T. cruzi* and *H. sapiens* superimposed on template structure. C) Focus on the active site. Top left: Template (orange) with WPF inhibitor (gray) and *T. cruzi* (cyan). Top right: *T. cruzi* and *H. sapiens* (green). Middle: Electrostatic potential (Kb.T.ec$^{-1}$) – blue: negative, red: positive. Left: *T. cruzi*, right: *H. sapiens*. Bottom: Hydrophobicity – orange: hydrophobic, blue: hydrophilic. Left: *T. cruzi*, right: *H. sapiens*. D) RMSD variation as the function of time (ns), relative to the initial structure. E) Superimposition of the initial structure (cyan) and most populated cluster (red) over 100ns period molecular dynamics.

**Figure 4:** A) 2,4-dienoyl-CoA reductase (NADPH) (EC: 1.3.1.34). *T. cruzi* model (cyan), superimposed on *E. coli* structure (PDB: 1PS9 - orange). B) Focus on the active site. Top: Template (orange) with FMN inhibitor (gray) and *T. cruzi* (cyan). Middle: Electrostatic potential (Kb.T.ec$^{-1}$) – blue: negative, red: positive. Bottom: Hydrophobicity – orange: hydrophobic, blue: hydrophilic. C) RMSD variation as the function of time (ns), relative to the initial structure. D) Superimposition of the initial structure (cyan) and most populated cluster (red) over 100ns period molecular dynamics.

**Figure 5:** A) Dimer structure of the Cysteine Synthase (EC: 2.5.1.47). *T. cruzi* model (cyan), superimposed on *L. donovani* structure (PDB: 3TBH - orange). Gray depicts the second monomer. B) Focus on only Chain A of *T. cruzi* superimposed on template structure. C) Focus on the active site. Top: Template (orange) and *T. cruzi* (cyan). Middle: Electrostatic potential (Kb.T.ec$^{-1}$) – blue: negative, red: positive. Bottom: Hydrophobicity – orange: hydrophobic, blue: hydrophilic. D) RMSD variation as the function of time (ns), relative to the initial structure. E) Superimposition of the initial structure (cyan) and most populated cluster (red) over 100ns period molecular dynamics.

**Figure 6:** A) Leishmanolysin (EC: 3.4.24.36). *T. cruzi* model (cyan), superimposed on *L. major* structure (PDB: 1LML - orange). B) Focus on the active site. Top: Template (orange) with Zinc Ion (gray) and *T. cruzi* (cyan). Middle: Electrostatic potential (Kb.T.ec$^{-1}$) – blue: negative, red: positive, gray: Zn ion. Bottom: Hydrophobicity – orange: hydrophobic, blue: hydrophilic. C) RMSD variation as the function of time (ns), relative to the initial structure. D) Superimposition of the initial structure (cyan) and most populated cluster (red) over 100ns period molecular dynamics.
Figure 1
Figure 2

A) cytoplasm
B) Extracellular

C) Structure

<table>
<thead>
<tr>
<th>Structure</th>
<th>D359</th>
<th>L361</th>
<th>F385</th>
<th>T392</th>
<th>K408</th>
<th>R441</th>
<th>L443</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3AR4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

D) 

E)
Article

Figure 4

A)

B)

C)

D)
Figure 5
Figure 6
Table 1: Grouping of sequences for their respective domains; Red, sequences identified as being fragment to set in stude.

<table>
<thead>
<tr>
<th>Group</th>
<th>TcruzDB ID</th>
<th>L</th>
<th>Description</th>
<th>EC</th>
<th>Pfam Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TC00.104705350763.19</td>
<td>646</td>
<td>H+ exporting ATPase</td>
<td>3.6.3.6</td>
<td>ATPase / Hydrolase</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053506649.20</td>
<td>898</td>
<td></td>
<td>3.6.3.6</td>
<td>ATPase / Hydrolase</td>
</tr>
<tr>
<td>2</td>
<td>TC00.1047053504507.5</td>
<td>388</td>
<td>trypanothione reductase</td>
<td>1.8.1.12</td>
<td>Pyridine nucleotide disulphide oxidoreductase</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053503555.30</td>
<td>492</td>
<td></td>
<td>1.8.1.12</td>
<td>Pyridine nucleotide disulphide oxidoreductase</td>
</tr>
<tr>
<td>3</td>
<td>TC00.1047053509941.100</td>
<td>693</td>
<td>2,4-dienoyl-CoA reductase (NADPH)</td>
<td>1.3.1.34</td>
<td>Oxidoreductase / Pyridine nucleotide - disulphide oxidoreductase</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053510303.210</td>
<td>717</td>
<td></td>
<td>1.3.1.34</td>
<td>flavino oxidoreductase / dependent oxidoreductase</td>
</tr>
<tr>
<td>4</td>
<td>TC00.1047053507297.10</td>
<td>247</td>
<td>cruzipain</td>
<td>3.4.22.51</td>
<td>Peptidase_C1</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053506529.550</td>
<td>322</td>
<td></td>
<td>3.4.22.51</td>
<td>Inhibitor_I29/ Peptidase_C1</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053507537.20</td>
<td>426</td>
<td></td>
<td>3.4.22.51</td>
<td>Inhibitor_I29/ Peptidase_C1</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053508595.50</td>
<td>521</td>
<td></td>
<td>3.4.22.51</td>
<td>Peptidase_C1/ DUF3586</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053509401.30</td>
<td>181</td>
<td></td>
<td>3.4.22.51</td>
<td>Peptidase_C1/ DUF3586</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053509429.320</td>
<td>467</td>
<td></td>
<td>3.4.22.51</td>
<td>Peptidase_C1/ DUF3586</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053507603.260</td>
<td>467</td>
<td></td>
<td>3.4.22.51</td>
<td>Peptidase_C1/ DUF3586</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053507603.270</td>
<td>467</td>
<td></td>
<td>3.4.22.51</td>
<td>Peptidase_C1/ DUF3586</td>
</tr>
<tr>
<td>5</td>
<td>TC00.1047053509005.50</td>
<td>467</td>
<td>triacylglycerol lipase</td>
<td>3.1.1.3</td>
<td>Lipase</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053505931.10</td>
<td>516</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053511211.90</td>
<td>567</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053510565.150</td>
<td>589</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053507623.110</td>
<td>568</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053508699.100</td>
<td>566</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053506699.90</td>
<td>621</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053509011.80</td>
<td>592</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053506587.100</td>
<td>566</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053509205.100</td>
<td>567</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053506163.10</td>
<td>567</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053506163.20</td>
<td>567</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td>6</td>
<td>TC00.1047053508813.40</td>
<td>567</td>
<td>leishmanolysin</td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053505965.10</td>
<td>567</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053506257.50</td>
<td>600</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053510899.10</td>
<td>566</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053505931.20</td>
<td>579</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053511203.10</td>
<td>566</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053504397.20</td>
<td>567</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053506921.10</td>
<td>621</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053508475.30</td>
<td>566</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053505615.10</td>
<td>592</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053508825.10</td>
<td>565</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053510873.20</td>
<td>565</td>
<td></td>
<td>3.4.24.36</td>
<td>Peptidase_M8</td>
</tr>
<tr>
<td>7</td>
<td>TC00.1047053507165.50</td>
<td>332</td>
<td>cysteine synthase A</td>
<td>2.5.1.47</td>
<td>Pyridoxal - phosphate dependent enzyme</td>
</tr>
<tr>
<td></td>
<td>TC00.1047053507793.20</td>
<td>332</td>
<td></td>
<td>2.5.1.47</td>
<td>Pyridoxal - phosphate dependent enzyme</td>
</tr>
</tbody>
</table>
Supplementar information

Fig S1: ATPase model superimposed on the 3B8C template. A) Information about template 3B8C from A.thaliana. B) Superimposed model ATPase of T.cruzi and PDB template structures. C) Varying amino acids of the active site. D) HHalign-Kbest alignment of ATPase and 3B8C used for 3D modeling.

A) 
PDB id: 3B8C
Function: Hydrolase
Organism: Arabidopsis thaliana
Resolution: 3.6 Å
Method of Resolution: X-RAY
EC number: 3.6.3.6
Length: 885
Coverage in T.cruzi sequence: 4-833

B) 
TCR  
RMSD = 1.18 Å

C) 
<table>
<thead>
<tr>
<th>Structure</th>
<th>TCR</th>
<th>D359</th>
<th>L361</th>
<th>F385</th>
<th>T392</th>
<th>K408</th>
<th>R441</th>
<th>L443</th>
</tr>
</thead>
<tbody>
<tr>
<td>3B8C</td>
<td>D372</td>
<td>I374</td>
<td>F400</td>
<td>T407</td>
<td>K423</td>
<td>R456</td>
<td>L458</td>
<td></td>
</tr>
</tbody>
</table>

D)
Suplementar information

Fig S2: HHalign-Kbest alignment of the ATPase sequence from T.cruzi and O.cuniculus (PDB: 3AR4) used for 3D modeling.
Fig S3: A) Fraction of residues matching Ramachandran $\phi$ / $\psi$ plot favorable regions (%). B) Corresponding MoldFold4 scores (large is best).

A) 

B) 

---

Supplemental information
Fig S4: HHalign-Kbest alignment of the Tripanothyone Reductase sequence from T. cruzi and T. brucei (PDB: 2WPF) used for 3D modeling

<table>
<thead>
<tr>
<th>TCR</th>
<th>2WPF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSKFDLVLGAGGGLMAATLVKVRAYVDVOTHGPPFSAALGGTVNVCVFW</td>
</tr>
<tr>
<td></td>
<td>.MSKFDLVLGAGGGLMAATLVKVRAYVDVOTHGPPFSAALGGTVNVCVFW</td>
</tr>
<tr>
<td></td>
<td>KKLMTGAOYEMHLRSEAGFGEFRILTALWKKLIAKVDEAVLNIKAMEEPRTEG</td>
</tr>
<tr>
<td></td>
<td>KKLMTGAOYEMHLRSEAGFGEFRILTALWKKLIAKVDEAVLNIKAMEEPRTEG</td>
</tr>
<tr>
<td></td>
<td>LEFFLGWGSLESKNVWRESADPASAVKLRTEEILLASGSRWMPNIPGIEHCISN</td>
</tr>
<tr>
<td></td>
<td>LEFFLGWGSLESKNVWRESADPASAVKLRTEEILLASGSRWMPNIPGIEHCISN</td>
</tr>
<tr>
<td></td>
<td>CAFYLPEPPRPRVTLTVGGFQSEVFAGTFKYKPGVGTCYAGMILKFEQNTIREETL</td>
</tr>
<tr>
<td></td>
<td>CAFYLPEPPRPRVTLTVGGFQSEVFAGTFKYKPGVGTCYAGMILKFEQNTIREETL</td>
</tr>
<tr>
<td></td>
<td>KOLTANGIOIITYKHPKVARLNDQGSSVTFEGKMDFLVMAIGRPRKFDLQGCA</td>
</tr>
<tr>
<td></td>
<td>KOLTANGIOIITYKHPKVARLNDQGSSVTFEGKMDFLVMAIGRPRKFDLQGCA</td>
</tr>
<tr>
<td></td>
<td>GVYYKQGGYVODVEYRSRNPHYAIGDTVRMTLPVAINEAAALVDTETPPRTKTDH</td>
</tr>
<tr>
<td></td>
<td>GVYYKQGGYVODVEYRSRNPHYAIGDTVRMTLPVAINEAAALVDTETPPRTKTDH</td>
</tr>
<tr>
<td></td>
<td>TVASAVFSPPPAPICTQLEEFVAVSLSFPLMHKVSGSKYFVAKITUHH</td>
</tr>
<tr>
<td></td>
<td>TVASAVFSPPPAPICTQLEEFVAVSLSFPLMHKVSGSKYFVAKITUHH</td>
</tr>
<tr>
<td></td>
<td>SDQTVLGVHLLGDNAPEIQOVICLKLNAKIDFYNITGVHPTSAEELCSMRTPSYY</td>
</tr>
<tr>
<td></td>
<td>SDQTVLGVHLLGDNAPEIQOVICLKLNAKIDFYNITGVHPTSAEELCSMRTPSYY</td>
</tr>
<tr>
<td></td>
<td>KGCKMEKLC..EQD5..</td>
</tr>
<tr>
<td>180</td>
<td></td>
</tr>
</tbody>
</table>

Supplementar information
Suplementar information

Fig S5: HHalign-Kbest alignment of the 2,4 Dienoyl CoA sequence from T.cruzi and E.coli (PDB: 1PS9) used for 3D modeling.
Suplementar information

Suplementar information

Suplementar information

Fig S8: HHalign-Kbest alignment of the Leishmanolysin sequence from T.cruzi and L.major (PDB: 1LML) used for 3D modeling.
Conclusions

CHAPITRE VI
Conclusions

According to the World Health Organization (WHO), 21 Latin American countries are endemic for Chagas disease, affecting 10 million people. Chagas’ disease, caused by the protozoan Trypanosoma cruzi (T. cruzi), is a parasitic illness endemic mostly in Latin America, and the identification of specific T. cruzi protein targets for drug development is desirable. In recent years, drug discovery for Chagas disease has evolved from limited compound testing in manual assays to sophisticated in vitro and in vivo assays. Drug discovery must consider that T. cruzi is a genetically diverse organism with phenotypically variations, and inhibitor susceptibility is a well-established phenotypic variation both in vitro and in vivo. In the next five years, 200,000 people living with Chagas disease will die from heart disease and related complications, and the BENEFIT trial [1] pursues a global initiative of diagnosis, treatment, and research.

In my thesis, by following a complex bioinformatics procedure, we have proposed structural models for five T. cruzi enzymes with ATPase, trypanothione reductase, 2,4-dienoyl-CoA reductase, cysteine synthase, and leishmanolysin functions. Based on sequence and active-site similarity with several organisms and notably H. sapiens, we show that these five protein targets, except cysteine synthase, are very attractive for drug development against T. cruzi [2].

Note that when constructed, the Homo sapiens analogs or homologs are not set to MD simulations. Ideally an apple-to-apple comparison with these analogues going through the same computational procedure should be performed. However our goal was not to perform a comparative study of structural models of T. cruzi and H. sapiens by MD simulations, but to identify within 41 protein targets therapeutic targets specific to T. cruzi. We found that among five enzymes, only two proteins with 2,4-dienoyl-CoA reductase and leishmanolysin activities are absent in H. Sapiens, and based on analysis of the electrostatic potential and hydrophobic/hydrophilic character of the residues of the active sites and the stability of all residues upon MD simulations, 4 proteins except cysteine synthase merit further investigation.

There are two on-going projects and perspectives related to the results of my thesis.

First, among our initial pool of 41 T. cruzi protein sequences, 8 sequences related to cruzipain (EC: 3.4.22.51) and one sequence related to triacylglycerol lipase (EC: 3.1.1.3) were eliminated after PFAM protein domain analysis because they include large domains
Conclusions

(> 250 amino acids) of unknown functions (e.g. DUF3586 in cruzipain) that lack any sequence homology with available PDB structures. We are currently attempting to model their 3D structures by meta-servers based on fold-recognition, I-Tasser [3] and even PepFold [4]. It is possible that simple biochemical and biophysical data obtained from size exclusion chromatography, FTIR and CD analyses may help distinguish one putative model among many others.

Second, four *T. cruzi* structures are currently used for in silico drug screening using multiple on-line servers and large datasets of small molecules. Drug development based on computer simulations is not straightforward, however. The results of a naïve virtual screening experiment based on the structure of *T. cruzi* and *H. Sapiens* Trypanothione reductase are presented Figure 27. In this experiment, we have screened 10,000 compounds taken randomly from a diverse library of compounds, as proposed by MTIOpenScreen [5]. Our aim was to sense how much the difference between the physicochemical properties of the active sites can be expected to impact small compound selection. However, as can be seen over the 1000 best compounds (results unchanged for the best 100 compounds), no significant difference in the distributions of some standard molecular properties can be observed. Although more descriptors should be analyzed, this poses the question of the selectivity of the interaction of the compounds for the *T. Cruzi* and *H. sapiens* enzymes, which can be essential to meet the requirement of low adverse effect of a drug. Looking more in details, the energies of the poses of the best ranked compounds are significantly low in both cases (-10.0 and -8.6 kcal/mol for *T. cruzi* and *H. Sapiens*, respectively), and suggest that low enough binding affinities could be expected for both *T. Cruzi* and *H. Sapiens*. Indeed, it could reveal very important to explore a much larger dynamic time-scale of the enzymes and enzymes/inhibitors using advanced simulation methods so as to evaluate the correct binding pose and time-consuming free energy procedures in order to predict accurate binding affinity that goes beyond simple scores.
Conclusions

Figure 27: Left panels: T. Cruzi. Right panels: H. Sapiens. Top panels): Distribution of some standard molecular descriptors for the 1,000 best compounds identified (generated using FAF-drugs [6]). Bottom panels: Poses generated for the best compound in green (Autodock-vina [7] energies are of -8.6 and 10.0 Kcal/mol for H. Sapiens and T. cruzi, respectively).
Conclusions


Annexe
Annexe

A Computational Methodology to Overcome Challenges Associated with the Search for Specific Targets to Develop Drugs against Leishmania major

Larissa Catharina Costa1,¶, Carlyle Ribeiro Lima2,4,¶, Ana Carolina Ramos Guimarães3, Marcelo Alves-Ferreira1, Pierre Tuffery4, Philippe Derreumaux5, Nicolas Carels1,*

1Laboratório de Modelagem de Sistemas Biológicos, National Institute of Science and Technology on Innovation on Neglected Diseases (INCT-IDN), Centro de Desenvolvimento Tecnológico em Saúde (CDTS), Fundação Oswaldo Cruz (Fiocruz), Rio de Janeiro, Brazil.

2Laboratoire de Biochimie Théorique, Institut de Biologie Physico-Chimique (UPR 9080), Centre National de la Recherche Scientifique (CNRS), Université Paris 7, Paris, France.

3Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (Fiocruz), Rio de Janeiro, Brazil.

4Molécules Thérapeutiques in silico (MTi), Institut National de la Santé et de la Recherche Médicale (INSERM), Université Paris Diderot (UMR-S 973), Paris, France.

5Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Université Paris Diderot, Sorbonne Paris Cité, IBPC, 13 Rue Pierre et Marie Curie, 75005 Paris, France.

¶The authors contributed equally to this work

*Corresponding author

E-mail: nicolas.carels@cdts.fiocruz.br
Annexe

Abstract

Background

Nearly 50% of the global rational drug design efforts are focused on developing inhibitors against enzymes suggested to be critical for metabolic processes in pathologies. In a previous study, a methodology was established to identify enzymes whose activities are essential to metabolic pathways within protozoa of the genera Trypanosoma and Leishmania that could serve as new therapeutic targets to fight infections and prevent disease. The occurrence of false positives due to available sequence annotations has been a recurring problem, resulting in significant errors. Here, we present an optimized approach for detecting specific enzymes involved in the host-parasite relationship to advance targeting for drug development.

Methods/Principal Findings

The methodology described here is based on traditional techniques of sequence homology comparison and on hidden Markov models; it integrates elements of enzymatic functionality, secondary and tertiary structures, protein domain architecture, as well as the metabolic environment. Applying this approach to Leishmania major, we identified sterol 24-C-methyl transferase, pyruvate phosphate dikinase, trypanothione synthetase and an RNA editing ligase as four enzymes strictly specific to parasites compared to humans. Each of these enzymes has been previously described in literature as specific to L. major providing validation of the methodology presented here.

Conclusions/Significance

The comparison of L. major and human has allowed the identification of specific enzymes of L. major that may serve as targets for drug development. Here, our multi-parameter methodology uncovered the potential of four previously described parasite-specific enzymes to serve as drug targets. We propose that these four targets should be addressed simultaneously by a cocktail of suitable inhibitors to disrupt parasite metabolism and reduce the likelihood of drug resistance.
Annexe

Author summary

Species of the genus Leishmania are obligate intracellular pathogens that cause Leishmaniasis by infecting cells of the immune system. Worldwide, nearly 350 million people in 88 countries are at risk of infection by the various species of Leishmania. Based on geographical distribution, Leishmania major is one of the main species of the genus that is transmitted by several species of sandflies vectors that belong to the Dipteran order and Phlebotomus genus. It is an excellent model for studies based on bioinformatics techniques due to its well annotated genome and the existing need for more effective treatments that would be accelerated by identifying high potential candidate enzymes for rational drug design. The aim of the work presented here is to propose a list of candidate enzymes based on a multi-parameter, bioinformatics methodology that pays special attention to identify of false positives. Although the number of potential targets for drug development is low with only four strict specific enzymes (sterol 24-C-methyl transferase, RNA editing ligase, pyruvate phosphate dikinase and trypanothione synthetase), we propose that a cocktail therapy including inhibitors for these four targets could stably control Leishmaniasis.

Introduction

The pharmaceutical industry has experienced a dramatic decrease in productivity between the 1980’s and 2010 (1–3) that is principally due to the cost burden of investing in the research and development (R&D) for new drugs, estimated to be $1.7 billion each (4). In response, the process of drug R&D is shifting from how these activities addressed healthcare priorities in the past to approaches that are being dominated by potential market value. Even if the situation reverted to the 1980s as defined by the number of new chemicals licensed by the FDA, concerns exist with regard to the future decision making behind appropriating R&D infrastructures that requires a new paradigm for the management of these activities to attend global needs (5).

One unattended global need is for new drugs against parasitic diseases endemic to tropical regions in developing countries. Due to their low market potential, R&D for drugs against these diseases is generally under prioritized or neglected by private companies and its realization depends on other means (6,7). Since alternative financial sources for research in neglected disease research is limited and highly divided among individual researchers, open science and data sharing have received a growing interest as a means of leveraging and combining available resources to accelerate drug discovery efforts (8). This community-based concept for a new drug discovery model led to the London Declaration on Neglected Tropical Diseases in 2012 for the control, elimination or eradication of neglected tropical diseases (http://unitingtocombatntds.org). In the
Annexe

spirit of this cooperation, we are contributing with the identification of four molecular targets for rational drug design to treat Leishmaniasis and describe our methodology for generating candidate list for other pathogens.

The first step of drug development relates to the identification of molecular targets relevant to disease control. While the biological targets for all or nearly all diseases were expected to be quickly revealed by genomic studies, the genomes from complex eukaryotes, including humans, have been shown to include more than 20,000 genes (9). Many of the associated proteins are potential targets for drug interventions to control human diseases. The most recent number of drug targets was estimated to be in the hundreds, based on analysis made before 2007 (10). However, the number of druggable proteins is substantially greater according to the DrugBank database website (http://www.drugbank.ca/). The current version of this database (5.0) contains 8206 drug entries that are linked to 4333 non-redundant protein sequences (i.e., drug target/enzyme/transporter/carrier). Since 4 to 5 years of research are required for each new target, the discovery of the whole set of these targets is not feasible if high throughput techniques are not utilized (11).

For rational drug design, there are basically four types of macromolecules with which therapeutic agents can interact: proteins, polysaccharides, lipids and nucleic acids. However, the difficulties in targeting with the last three molecules have encouraged more research on protein inhibitors (10). More precisely, enzymes are most frequently candidates, accounting for nearly 50% of the total pharmacological targets being addressed worldwide (12). The specific inactivation of a pathogen enzyme without affecting any human enzymes would provide a safe approach to control the pathogen. The discovery of drugs based on the in silico docking of inhibitors in models of the three-dimensional (3D) structures of protein targets has proven to be of great value in the process of drug screening and has effectively contributed to conserving resources in the area of drug discovery. The use of this information may allow substantial saving of cost and time involved in the process of drug release (13,14).

Potential proteins for screening include those originating evolutionarily from independent sources or diverging from an ancestral sequence over the course of a billion years, which separates the common origins of unicellular and simple eukaryotes from phyto- and metazoan. Both mechanisms could have given rise to enzymatic functions specific to parasitic pathogens in comparison to their hosts (15). Similar mechanisms could generate analogous enzymes in pathogens and hosts that can also provide potential molecular targets due to variations in their enzymatic sites that can be distinguished for inactivation by a given inhibitor (14). By definition, analogous enzymes result from the convergent evolution of independent proteins rather than originating from
common ancestral proteins (homologous), which allows for the same function, but with differences in their primary, secondary and tertiary structures (13).

A previous study established a methodology to identify enzymatic activities in essential metabolic pathways that can serve as new therapeutic targets to fight infectious and parasitic diseases caused by protozoa of the Trypanosoma and Leishmania genera (16). This methodology includes: (i) a massive computational search for genes that encode enzymes that are specific to parasites and are not encoded in human DNA that includes analogous enzymes, and (ii) the in silico evaluation of the biological significance of the potential drug targets within the context of the entire metabolism of the pathogens under consideration. The resulting pipeline, called AnEnPi (Analogous Enzymes Pipeline), identifies, annotate and compares specific and analogous enzymes using an input of the protein sequences stored in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (http://www.genome.jp/kegg/). AnEnPi implements algorithms that can perform the following tasks: (i) the clustering of protein sequences by homology using BLAST (17); (ii) the classification of homologous sequence pairs in analogous enzymes when they have the same enzymatic function (EC number), but their homology score is not significant (18); (iii) the detection of specific and analogous enzymes (19); (iv) the annotation of protein functions using BLAST or HMMER; and (v) the generation of metabolic maps using the tools provided by KEGG (http://www.genome.jp/kegg/tool/map_pathway1.html). The reconstruction of metabolic pathways in parasites based on the metabolic maps provided by KEGG aims to identify enzymatic activities that are essential for a parasite and that can be considered promising targets for drug development (20–22).

As a pipeline, AnEnPi classifies enzymes into homologous, analogous and specific by only considering the existing functional annotations of the enzymes (EC numbers) in the two organisms being compared. Enzymes were designated homologous when the similarity score obtained with their primary sequences comparison was ≥120 (using BLASTp algorithm), analogous when their similarity score was ≤120. A homologous pairs could match enzyme with different EC numbers and consequently named as functional specific. Under that classification, two analogous enzymes are associated to the same EC number, but their homology score is ≤120. In contrast, when an enzyme had all its similarity scores ≤120 and did not share any EC number with other enzymes, it was designated as strict specific.

Any error associated with a classification process can be qualified as either: (i) false positives when classified as specific or analogous and at least one homolog exists in the H. sapiens genome; (ii) false negatives when the classification of a homologous pair is erroneous and should be
Annexe

classified as specific or analogous because of an incorrect annotation; (iii) true positives when a pair of specific or analogous genes is correctly classified; and (iv) true negatives when a pair of homologous genes, indeed, does not refer to specific or analogous genes.

The occurrence of false positives based on available sequence annotations can result in significant errors and has been a recurring problem in the previously described methodology. Often, incomplete annotations are responsible for generating false positives of specific or analogous enzymes. Since annotation is a dynamic process that improves with time, at any given moment, a pair of enzyme-encoding genes can be incorrectly classified as specific simply because the homologous counterpart in one genome of the pair has not yet been annotated.

Here, this report focuses on an evaluation of the false positives produced by AnEnPi because (i) in its current iteration, AnEnPi does not have any process of evaluating the generation of false positives and (ii) this error component is critical for decisions on the investment of limited resources on a putative specific or analogous enzyme for drug development. The existence of false positives during target identification could invalidate years of research into drug development that could be extremely difficult to recuperate. Errors resulting from incomplete annotations are considered here, not those from mistakes in genome assemblies or EC number designations, and we describe a method to track false positives in the AnEnPi output to improve the identification of enzymes specific to L. major compared to H. sapiens.

**Materials and Methods**

We searched for false positives in a list of enzyme sequences obtained from AnEnPi (16) whose functions were annotated as specific for L. major compared to H. sapiens. Here by enzymes specific of L. major, we meant enzymes whose reactions (EC number) were found to be catalyzed in L. major but not in H. sapiens. Thus, an enzyme was considered specific in L. major when its EC number was not found in H. sapiens. The gene list included 67 sequences from L. major with 42 being associated to a respective EC number that was classified as specific to L. major compared to H. sapiens by AnEnPi.

We could distinguish several components to the process of false positive diagnostics. (i) The amino acid sequence comparison between L. major and human proteins allowed us to quickly identify most of the human proteins that deserved attention as potential false positives. (ii) The comparison of L. major proteins with human DNA was straightforward and gave the sequence coordinates where homologous sequences could be found in the human genome. (iii) The comparison of human proteins that were homologous to a L. major protein to human chromosomal
Annexe

DNA allowed the delineation of its gene structure (exons and introns) corresponding to the human query proteins. (iv) The comparison of the L. major proteins that had a homologous hit within the human genomic sequence without a corresponding human protein in the Ensembl list (23) indicated a potentially missing human gene annotation or a possible pseudogene. (v) The consistency of homology detection could be analyzed by performing multiple comparisons to determine if the same matches were obtained from protein to protein and protein to DNA using parasite to human and human to parasite queries.

To determine whether the classification of specific enzymes might produce false positives, we first searched for homology between the proteins of L. major and the DNA sequence of the human genome. To this end, we employed the sequences from the putative L. major-specific proteins available from TriTrypDB (release 6.0 – September 2013; (24)) and compared them to the human genome sequence (Ensembl, release 74 – November 2013) using tBLASTn (25). We considered a pair as a consistent homologous hit when the E-value of a tBLASTn alignment was ≤10^{-4} and its score value ≥120. When consistent hits were obtained, the genomic coordinates from the tBLASTn output were recovered and compared to the genomic coordinates of encoded protein sequences available from Ensembl (release 74 – November 2013) to determine whether a protein annotation might exist for the genomic region corresponding to the tBLASTn hit. When such an Ensembl protein annotations did exist, it was checked for consistency with the gene model available from NCBI CCDS (26–28) and, if consistent, it (them) was (were) assigned to the considered genomic region where the homology with L. major protein was obtained. We also confirmed the homologous matches between L. major protein queries and the human protein sequences (subject) corresponding to their genomic hits with BLASTp (PSI-BLAST), (17) (see a detailed description in the supplementary file S1 and the corresponding table in supplementary file S2).

In cases where a genomic hit did not match the position of any human protein in the Ensembl list, it was considered an indication that the human gene had not received a designation during the annotation process of Ensembl. The absence of an annotation was addressed by translation of that homologous stretch of the human genome (subject) into amino acids followed by a homology search (BLASTp) with the non-redundant (nr) section of GenBank (release 201.0 – April 2014). This approach allowed the recovery of the corresponding annotation, associated CDS and its intron/exon gene structure model from the significant homologies.

Considering instances of enzymatic functional specificity between two homologous proteins of L. major and H. sapiens where the proteins were annotated to have differences in their enzymatic functionality based on the assigned EC numbers, we distinguish the following categories: (i) False
Annexe

positives for enzymatic specificity when the identity of the enzymatic function in both proteins cannot be confirmed. This conservative position was taken because when the members of a homologous pair have the same function (or EC number), the potential of negative collateral effects on the host of an inhibitor against the parasite enzyme is greater if both host and parasite enzymes are orthologues. (ii) True positives for enzymatic functional specificity in cases where the difference in enzymatic function of the homologous proteins could be confirmed. When both genes in the pair are paralogues rather than orthologues, they have an increased likelihood of having some sort of functional enzymatic specificity (15). However, an inhibitor designed against the L. major target may still affect the paralogous host counterpart with a lower rate, which requires a greater allocation of resources to confirm. (iii) Uncertainty with the annotations. In these cases, we investigated the protein’s name and function in UniProtKB/SwissProt (29) and attempted to verify whether the function described in this database matched that of TriTrypDB. If the function of the enzymes under comparison was synonymous in both organisms, the L. major protein was diagnosed as a false positive for functional enzymatic specificity with regard to its human counterpart; otherwise, functional specificity was assigned to the L. major enzyme.

In another attempt to discriminate false positives, we also mapped protein domains using the Pfam database (30), the Conserved Domains Database (CDD) (31) and as analysis of secondary structures (2D) using PRALINE (32) for the conserved regions of H. sapiens and L. major homologous pairs.

Remote homology detection with hidden Markov models

In addition to BLAST, we used HHpred (http://toolkit.tuebingen.mpg.de/hhpred; (33,34)) and HHsearch (http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::HHSearch; (35)), which implement hidden Markov models (HMM) for a pairwise comparison of the profiles from sequence alignments in databases like Pfam or SMART. HHpred detects the homologous sequences in humans with higher sensitivity than BLAST. HHpred was executed with the default options, i.e., local alignment and the scoring of secondary structure similarity, and HHsearch was executed with pdb70 (as the reference database), HHblits (as the alignment generation method; (36)), 85% (as sequence coverage), 10-6 (P-value), and 10 (as maximal hit number). The sequences resulting from the first BLAST screening were further analyzed through HMM HHpred and HHsearch to identify the possible remote sequence homologies in the human proteome that could result in false positives for the L. major-specific enzymes. We considered proteins to be homologous pairs when their identities were larger than 35% over a sequence coverage of at least 85% by HMM.
**Metabolic pathway significance**

In order to clarify the relevance of putative enzyme targets of *L. major* in terms of the metabolic impairment that results from their inactivation, we searched their respective EC numbers in the KEGG pathway database (http://www.genome.jp/kegg/pathway.html). Then, we searched their metabolic pathway insertion in order to diagnose whether their inactivation may potentially affect *L. major* without having deleterious consequence on the *H. sapiens* metabolism and consequently be used as targets for drug development.

**Homologous comparison of *L. major* specific enzymes with other parasitic protozoa**

Because an enzyme that is specific of *L. major* compared to *H. sapiens* could also have homologous pairs in other human parasites, we searched for sequence homologies between confirmed *L. major* - true positives proteins (proteins without any homologies with the human genome - *L. major* protein sequences compared them to the human genome sequence using tBLASTn) and proteins from the three protozoan parasites: *Trypanosoma cruzi*, *Trypanosoma brucei* and *Plasmodium falciparum*. The protein sequences from *T. cruzi* (TriTrypDB-6.0_Tcruzi.fasta contains sequences from diverse strains identifies as TcruziCLBrener, TcruziCLBrenerEsmeraldo-like, TcruziCLBrenerNon-Esmerald and TcruziSylvioX10-1) and *T. brucei* (TriTrypDB-6.0_Tbrucei.fasta includes sequences from diverse strains identifies as TbruceigambienseDAL972, TbruceiLister427 and TbruceiTREU927 strains) were retrieved from the TriTrypDB database (release 6.0 – 11/09/2013). The protein sequences from *P. falciparum* were retrieved from the Broad Institute (*P. falciparum* Dd2 nucleus - 09/04/2014).

**Results**

**Enzyme classification**

From the 67 protein sequences and 42 corresponding enzymatic activities that were identified as specific by AnEnPi, 15 sequences from *L. major* produced 55 homologous hits with the human genome, leaving 52 *L. major* proteins without any homologies with the human genome (true positives). Among the set of 15 proteins, (i) LmjF.18.1510 and LmjF.18.1520 were 99% identical and had the same EC number, but localized to different genomic locations. Similarly, LmjF.24.2030 and LmjF.27.2440 were homologous with the same EC number. Overall, among the 15 unspecific proteins, we found 13 different enzymatic activities (EC:1.1.1.100, 1.1.1.60, 1.3.1.71, 1.4.1.4,
Annexe
1.5.1.33, 1.8.1.12, 2.1.1.64, 3.1.4.46, 3.4.21.83, 3.5.5.1, 3.6.3.6, 4.2.1.70, 4.2.3.1). Within these 13 enzymatic activities, three (EC:3.1.4.46, 3.6.3.6, 4.2.1.70) were shared with the set of 32 EC numbers that corresponded to those of the remaining 52 specific proteins. This left a subset of 47 specific proteins with 29 EC numbers for specific enzymatic activities. Because LmjF.28.2100 (EC:3.1.4.46), LmjF.18.0506 (EC:3.6.3.6), LmjF.19.1020, LmjF.26.0420, LmjF.36.1660 (EC:4.2.1.70) were determined to be homologous unspecific enzymes, they were removed from the list of strictly specific enzymes (Fig 1).

Fig 1. Venn diagram of L. major proteins according to enzymatic activities and homologies with the human genome. The numbers for the circles and rounded rectangles represent the enzymatic activities in each category. The numbers in parentheses represent the protein number per category. Squares symbolize clusters of proteins per enzymatic activity with the number of clusters given inside the box along with the total number of proteins involved. Hexagons symbolize unique proteins according to enzymatic activity with their numbers inside.

Two or more enzymes encoded by different genes in an individual genome that share an identical enzymatic function (EC number) can be produced by intragenomic paralogous or analogous proteins. Enzyme redundancy is a major hurdle for drug development because protein isoforms for an identical enzymatic function can react differently to a drug and provide resistance. Therefore, we searched for enzyme redundancy in L. major through a sequence similarity search (BLASTp) and determined the homology status (paralogues or analogues) according to the homology score using <120 for analogous or ≥120 for paralogous, as described by Galperin et al. (18).

Among the 67 putative specific protein sequences released by AnEnPi, 38 proteins were classified according to their EC numbers into 13 homologous clusters, with seven clusters containing two proteins, three clusters containing three proteins, one cluster containing four proteins, one cluster containing five proteins and one cluster containing six proteins (see Table 1 in supplementary file S3). The 29 protein sequences of L. major that did not form intragenomic homologous clusters were divided into functional specific (LmjF.05.0350, LmjF.09.0770, LmjF.14.0350, LmjF.23.0270, LmjF.26.2280, LmjF.28.2910, LmjF.30.0180, LmjF.33.0680, LmjF.35.4250) and strictly specific enzymes (LmjF.07.0260, LmjF.07.0270, LmjF.09.1040, LmjF.11.1000, LmjF.14.0460, LmjF.16.0530, LmjF.17.0140, LmjF.17.1160, LmjF.17.1360, LmjF.18.0200, LmjF.26.0830, LmjF.26.2610, LmjF.29.2800, LmjF.30.0610, LmjF.30.3080, LmjF.31.0010, LmjF.34.0070, LmjF.34.3250, LmjF.36.3590, LmjF.36.4640).
A comparison of the gene structure obtained by matching Ensembl protein sequences and their human DNA equivalent allowed us to identify those enzymes that could represent true homologies between L. major and humans. Among the 15 unspecific L. major proteins, nine were unambiguously annotated with different EC numbers and were designated to be associated with different enzymatic reactions in L. major and humans. As a result, we considered these nine proteins to be functionally specific enzymes (see Table 2 in supplementary file S3 for a detailed description).

Despite differences in EC annotations, we did not find any obvious evidence for functional specificity in the homologous protein pairs from L. major and H. sapiens by a cursory examination of their 2D profile comparisons due to their high level of similarity (see supplementary file S3 for a detailed description).

Remote homology detection with hidden Markov models

We used HHsearch to identify templates that could be used to generate higher quality 3D models. The use of both HHpred and HHsearch improved the models by 50% to 100% according to the alignment quality and was approximately 4.2 times more sensitive compared to BLAST according to remote homology detection (below the twilight zone of 35% amino acid sequence identity). The Venn diagram in Fig. 1 shows that 11 sequences corresponding to homologous sequences between L. major and H. sapiens were considered, a priori, as being functionally specific for L. major since they were annotated with different EC numbers compared to the human sequences. A BLAST search also revealed 47 sequences that were reported to be strictly specific for L. major, while nine were at the intersection between functionally specific and strictly specific for L. major. When the HMM was performed on these three sequence sets, we found seven homologies in the set of sequences that were considered to be strictly specific for L. major according to the BLAST using the criteria of a minimum coverage of 85% and a minimum identity rate of 37%. Significant HMM homologies with the H. sapiens counterparts were identified for LmjF:07.0270, 14.0180, 16.0530, 17.0140, 18.0200, 33.2540 and 36.3590, which were considered false positives for enzymes that were strictly specific to L. major since they have homologous pairs in H. sapiens with the same EC numbers as L. major.

Specific enzyme functions in L. major

Most of the 40 protein sequences associated to strictly specific enzymes of L. major after HMM filtering (see enzymatic reactions in the supplementary file S4) were associated to 23 ECs number (EC:1.1.1.3, 1.1.3.8, 1.11.1.11, 1.14.13.1, 1.7.1.1, 2.1.1.14, 2.1.1.41, 2.3.1.158, 2.7.1.39, 2.7.7.64, 2.7.9.1, 3.1.1.73, 3.1.30.1, 3.2.1.26, 3.2.2.1, 3.4.24.36, 4.1.1.74, 4.1.99.3, 4.4.1.8, 6.2.1.12,
Annexe

6.3.1.1, 6.3.1.9, 6.5.1.3) available from KEGG. Among these 40 strictly specific enzymes of L. major ten were distributed in eight EC numbers without any previous report in Tritryps in the literature: EC:1.1.1.3 – LmjF.07.0260; EC:1.1.3.8 – LmjF.17.1360; EC:1.14.13.1 – LmjF.26.2610; EC:1.7.1.1 – LmjF.30.0610; EC:2.1.1.14 – LmjF.31.0010; EC:2.3.1.158 – LmjF.09.1040; EC:3.1.1.73 – LmjF.36.4640; EC:6.2.1.12 – LmjF.19.0985, LmjF.19.0995 and LmjF.19.1005. The remaining 30 strictly specific enzymes had at least one reference associated to some species of Leishmania or Trypanosoma genus.

For proteins of L. major that had no EC number assignment in KEGG, we used the annotations from TriTrypDB, UniProtKB and GeneDB as describe in Table S4. Considering the metabolic pathways in which the 23 EC numbers were located, we found that one was associated with energy metabolism (EC:2.7.9.1), one to steroids synthesis (EC:2.1.1.41), seven to amino acids metabolism (EC:1.1.1.3, 2.1.1.14, 2.7.1.39, 4.1.1.74, 4.4.1.8, 6.2.1.12, 6.3.1.1), two to carbohydrate metabolism (EC:3.1.1.73, 3.2.1.26), two to ascorbate metabolism (EC:1.1.3.8, 1.11.1.11), one belongs to both carbohydrate and ascorbate metabolisms (EC:2.7.7.64), two to nucleotide metabolism (EC:3.2.2.1, 4.1.99.3), two to nucleic acid metabolism (EC:3.1.30.1, 6.5.1.3), one to nitrogen metabolism (EC:1.7.1.1), one to the pathways of dioxin, naphthalene, polycyclic aromatic hydrocarbon degradation (EC:1.14.13.1), one to glycolipid metabolism (EC:2.3.1.158) and one to glutathione metabolism (EC:6.3.1.9). The last EC number (3.4.24.36) was correlated to six sequences with the protease activity of the highly abundant zinc metallopeptidase, which are named leishmanolysin or GP63 and have been related to a myriad of functions involving host-parasite interactions (37).

Most of the 40 specific enzymes were found in pathways with alternative routes and, therefore, are not expected to be essential. Only eight specific enzymes corresponding to four ECs (EC:2.1.1.41 - LmjF.36.2380 and LmjF.36.2390; EC:2.7.9.1 - LmjF.11.1000; EC:6.3.1.9 - LmjF.23.0460, LmjF.27.1870 and LmjF.36.4300; EC:6.5.1.3 – LmjF.01.0590 and LmjF.20.1730) could serve as potential target candidates for drug development.

**Homologous comparison of L. major specific enzymes with other parasitic protozoa**

To speed up the process of identifying a specific protein target for drug development using other human parasitic protozoa, we searched for homologous proteins to the 52 L. major true positives enzymes in T. cruzi, T. brucei and P. falciparum. In the comparison between L. major (query) and T. cruzi (subject: 34,187 proteins), we found that 17 L. major proteins were associated with 47 T. cruzi homologous pairs whose identity level (Id) was greater than 60%, 33 L. major
proteins had the Id < 60% and two L. major proteins (LmjF.23.0870 - EC:3.2.1.26 and LmjF.31.0010 - EC:2.1.1.14) did not match any T. cruzi proteins.

In the comparison of L. major with T. brucei (29,302 proteins), we found that ten L. major proteins were associated with 29 T. brucei homologous pairs (Id ≥ 60) and that the L. major proteins LmjF.23.0870 (EC:3.2.1.26) and LmjF.07.0260 (EC:1.1.1.14) did not match any T. brucei protein (Id < 60).

When comparing L. major with P. falciparum (5,139 proteins), we found that 3 L. major proteins, LmjF.10.0460 (EC:3.4.24.36), LmjF.10.0465 (EC:3.4.24.36) and LmjF.18.0560 (EC:3.6.3.6) were associated with 2 P. falciparum homologous (Id ≥ 60%) pairs PFDG_02171 and PFDG_00701, while 35 L. major proteins had an Id < 60. The L. major proteins LmjF.09.1040 (EC:2.3.1.158), LmjF.10.0470 (EC:3.4.24.36), LmjF.10.0480 (EC:3.4.24.36), LmjF.14.0180 (EC:3.4.17.19), LmjF.17.1360 (EC:1.1.3.8), LmjF.19.1020 (EC:4.2.1.70), LmjF.23.0460 (EC:6.3.1.9), LmjF.23.0870 (EC:3.2.1.26), LmjF.26.2610 (EC:1.14.13.1), LmjF.30.0308 (EC:2.7.1.39), LmjF.31.0010 (EC:2.1.1.14), LmjF.31.2000 (EC:3.4.24.36), LmjF.35.0640 (EC:3.2.1.26) and LmjF.36.4300 (EC:6.3.1.9) did not match any P. falciparum protein (supplementary file S5).

The absence of homology for LmjF.23.0870 (EC:3.2.1.26) in T. cruzi, T. brucei and P. falciparum suggests that this enzyme is specifically restricted to Leishmania spp. within Tritryps. Actually, in KEGG EC:3.2.1.26 was identified in all Leishmania spp. but not in any one of T. cruzi or T. brucei.

**Discussion**

In this study, we elaborated a process for filtering out the false positive annotations of enzymatic specificity in host-parasite interactions. The identification of specific protein targets in parasites (here, L. major) compared to their host (here, H. sapiens) is important for the development of drugs with the least amount of negative collateral effects for the host as possible. The procedure presented here is systematic, compatible with automation and suitable for host-parasite interactions involving a lower eukaryote as the parasite and a higher eukaryote as the host. Considering that the targets identified here have been previously identified suggests that the results are valid and increases the confidence that the methodology proposed can be applied to other host-parasite systems with a lower level of associated knowledge.

We analyzed 67 protein sequences that correspond to 42 enzymatic activities previously
Annexe

classified as specific for L. major compared to H. sapiens by AnEnPi (16). A careful homology comparison of these sequences with Ensembl using BLAST identified 15 significant homologous pairs between L. major and H. sapiens that were associated with different EC numbers in L. major and H. sapiens, suggesting that they should be classified as functionally specific for L. major compared to H. sapiens. Of course, the classification into enzymes with functional specificity is completely dependent on the EC number annotation and may result in trivial errors, as is the case for the ATPases that may carry H+ or Na+ cations through cell membranes. The two ATPases carrying H+ or Na+ are homologous, and the difference in their EC numbers is associated with the transported cations (the reaction and not the enzymatic activity), which are not associated with the ATPase activity itself when the substrate is ATP. Thus, care must be taken when interpreting the EC numbers. In any event, we only found minute differences when considering the profiles of 2D structures between the sequence members of the BLAST homologous pairs.

Interestingly, the enzymes in this study that were considered to be functionally specific for L. major were not regarded as homologous by HHsearch, which indeed supports their classification as functionally specific for L. major. However, according to a conservative approach, we preferred to disregard the group of L. major enzymes for further investigations in the context of drug development. Among the 15 putative functionally specific enzymes, the EC numbers of nine L. major proteins only differed from those of H. sapiens in the fourth digit. Thus, these nine cases would be better referred to as functional specificities. When comparing the 2D alignments of these nine homologous pairs, we only found slight differences, and it is difficult to assess whether a drug for these L. major targets would have some kind of unwanted interaction with the human form that could result in negative side effects for the patients.

The Ensembl database involves processes of automatic gene annotation that are eventually manually curated, as it is the case for Havana (23,38). The AnEnPi pipeline is completely automatic and classifies enzymes as specific, homologous or analogous according to their homology and enzymatic activities. Obviously, incomplete enzymatic annotations enable AnEnPi to provide non-optimized classifications. The purpose of this study was to detect these cases and provide more knowledge to improve the general performance of the automatic classifier AnEnPi.

Although sequencing methods have been producing large amounts of data, the genome assemblies based on these data may be incomplete or approximate (39). Low-quality assemblies result in low-quality annotations (40,41) and promote both over- and underestimations of the numbers of genes (42,43). Considering the subject organism in a comparison, the incomplete assembly of its genome may lead to gene loss and thus generate a virtual specific gene in the query.
Annexe

genome. It is possible to check the consistency of an event that is observed, but it is almost impossible to check the consistency of an event that is not observed. Thus, in this study, we did not take misassemblies into account as a source of false positives. Only an in vitro analysis of our set of putative true positives can address this question.

For example, the existence of pseudogenes can be virtual or real, depending on whether they are the consequence of some mistake in the sequencing (induced frame shifts and premature stop codon) or assembling (gene cut) processes that lead to inaccurate annotations (44–48). Indeed, many true pseudogenes have some form of biological activity, and, therefore, their accurate annotation is potentially important to drug development (48). We observed only one case of a human pseudogene associated with a protein of L. major. This pseudogene was associated with four homologous human proteins retrieved from nr (GenBank) by BLASTp whose human genomic coordinates did not match any of the protein sequences from Ensembl.

The homologies obtained by implementing HHpred and HHsearch have a higher confidence level compared to BLAST because HMMs considers the similarity in the amino acid sequence pairs from L. major query and human sequence, as well as the 2D and 3D structures of each sequence. This approach tends to ensure a higher level of confidence about the conclusion one can draw from remote homologies between a query and a subject in sequence pair comparisons since 3D alignment is known to be more informative than primary sequence alignment (49). HHpred performs a profile-profile search through the HMM algorithm in the InterPro database (50), which was specifically prepared for that purpose (36). InterPro is a database that integrates predictive information about a protein’s function from a number of partner resources, providing an overview of the families to which a protein belongs and the domains and sites it contains. The information space of InterPro covers more than 75% of the UniprotKB sequences (51). The contents of InterPro consist of diagnostic signatures and the proteins that they significantly match. The signatures consist of models (simple types, such as the regular or more complex expression patterns in HMM) that describe protein families, domains or sites. The models are constructed from the amino acid sequences of known families or domains, and they are subsequently used to search for and classify unknown sequences. Each of the member databases (CATH/Gene3D, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY, and TIGRFAMs) of InterPro contribute to a different niche of specific sub-family classifications (52). The version of HHsearch hosted on the Mobyle server supports searches of the crystallographic models of proteins in the PDB database. HHpred and HHsearch are intrinsically complementary to BLAST, but significantly, slower which justifies their use as filters downstream of BLAST.
Enzymatic activity can play a key role in a metabolic pathway. Every pathway is associated with a group of proteins that is a list of enzymes with their EC numbers (20). Therefore, a metabolic pathway is an oriented graph whose vertices are characterized by an enzyme and its associated chemical reaction (53). A key enzyme is an enzyme that is contained in a path, lacks an alternative option, and plays an important role in the survival of the organism. Consequently, the inhibition of the enzyme necessarily results in the inhibition of the corresponding pathway, and the organism debilitation. However, if a pathway that does not play an important role in the survival of the organism is targeted, one cannot inhibit the pathway, and the organism can continue normally. Thus, the key feature for exploring a metabolic pathway is to identify the enzymes that play a key role in a pathway and determine whether they are essential for the survival of the parasite, i.e., whether its inhibition is deleterious to that organism.

In the context of drug development, the ideal drug should be the one that could inhibit the activity of a specific enzyme in an essential pathway of the target organism with no alternative route. Strict specificity, i.e., the existence of a given enzyme in the parasite and not in the host, is obviously the best situation. However, functional specificity, where substrate specificity can be identified for the parasite form compared to the host form, can be contemplated as well. Actually, EC:1.3.1.71 (L. major), which is associated with the steroid biosynthesis pathway, is an example of functional specific enzyme homologous to the human enzyme EC:1.3.1.70. EC:1.3.1.71 could be explored for drug development of antifungal and antitrypanosomal agents (54). Xu et al. (2014) showed that the knockdown of the gene encoding the 14-alpha-demethylase enzyme (EC:1.14.13.70), which is upstream EC:1.3.1.71, in L. major contributed initially to a dramatic change in the profile of the lipid composition in amastigotes and to a virulence attenuation, which was reversed after a few weeks. Since the activity of EC:1.14.13.70 is common to the route of cholesterol and ergosterol syntheses in mammals and Leishmania spp., respectively, it is possible that L. major got around the loss of EC:1.14.13.70 function in the knockdown parasites by using sterols from the host. This escape strategy would be unlikely to occur if EC:1.3.1.71 was inhibited rather than the 14-alpha-demethylase enzyme activity in Leishmania spp. since EC:1.3.1.71 is inserted in the pathway of ergosterol synthesis that is specific to Tritryps. We did not find other potential reliable functionally specific cases in this study because additional alternative metabolite routes may be available to the parasite to sustain the function that was inhibited by drug treatment (55). More recently, McCall et al. (2015) has demonstrated using knockout techniques and inhibitors the importance of 14-alpha-demethylase enzyme in L. donovani (56).

Considering strictly specific enzymes, they can be divided in targets with an a priori lower and a higher potential for drug development. We briefly review, hereafter, the strictly specific
enzymes that have a lower potential for drug development because of the existence of possible alternative pathways or other potential hurdle.

The leishmanolysin shown to be a potent virulence factor that can influence the host mammalian innate immune system at cellular and molecular levels, and have been described in different species of Leishmania and Trypanosoma (37,57). Therefore, the search for inhibitors of this protease has been discussed and encouraged for the development of new therapeutic strategies against Leishmaniasis (58).

Homoserine dehydrogenase (EC:1.1.1.3 – LmjF.07.0260) is a strictly specific enzyme without description in Leishmania spp. However, recently it has been shown that this enzyme is also present in the pathogenic fungus Candida albicans and its importance in amino acid metabolism, protein, and cell adhesion has been demonstrated by knockout genomic experiments (59). In addition, Zhan et al. (2014) obtained the 3D structure of homoserine dehydrogenase from Mycobacterium leprae by comparative modeling and identified a number of compounds that docked to the binding sites of the substrate enzyme and of its NADH cofactor. The structure of these compounds may be used as scaffolds for the development of new inhibitors of this enzyme in Leishmania spp (60).

Among strictly specific enzymes involved in amino acid metabolism, two (EC:2.1.1.14, 6.2.1.12) do not have any reference associated with trypanosomatids in the literature, four (EC:2.7.1.39, 4.1.1.74, 4.4.1.8, 6.3.1.1) were described in other trypanosomatids and only one (EC:4.4.1.8) was identified in L. major. Homoserine kinase (EC:2.7.1.39) seems to be important for the growth of the promastigote form of T. brucei in the insect vector due to its participation in the threonine synthesis pathway (61). Indolepyruvate decarboxylase (EC:4.1.1.74) has been described in Phytonomas serpens, a trypanosomatid that infects plants, as taking part in the NADH reoxidation process of energy metabolism but its role in tryptophan metabolism was not determined (62). Williams et al. (2009) demonstrated the existence of two cysteine synthesis pathways in L. major: the de novo biosynthetic pathway and the reverse transsulfuration (RTS) pathway (63). The enzyme cystathionine β-lyase (EC:4.4.1.8) catalyzes a RTS pathway step, however these authors did not characterize this enzyme kinetically. A review of pyridoxal phosphate (PLP) dependent enzymes in protozoan parasites discussed their potential to be new therapeutic targets (64). More recently the 3D structure of L. major cystathionine β-lyase was solved by homology modelling (65). Asparagine synthetase (EC:6.3.1.1) has been described in several trypanosomatids (66–68), however its potential as a drug target is controversial. In Leishmania donovani it was demonstrated that this enzyme plays a key role in the viability of the pathogen (67). On the other hand, deletion of the gene coding for the asparagine synthetase in Leishmania infantum did not affect neither the
parasite growth nor its infectivity (68).

The invertase or β-fructofuranosidase (EC:3.2.1.26) is a strictly specific enzyme identified in L. major that has the function of performing the hydrolysis of sucrose releasing fructose and glucose. This enzyme has recently been described in different species of Leishmania and has been demonstrated to be intra- and extracellularly expressed (69,70).

Three strictly specific enzymes related to the metabolism of ascorbate were also identified. It is known that trypanosomatids are able to synthesize ascorbate in glycosomes unlike humans that are dependent on their diet to obtain vitamin C from external sources (71). Two of these enzymes have already been characterized in L. major: The ascorbate peroxidase (EC:1.11.1.11) and the UDP-sugar pyrophosphorylase (EC:2.7.7.64) (72–74); the latter also plays an important role in the carbohydrate metabolism (73,74). The third enzyme (EC:1.1.3.8) has not been described in L. major, however another galactonolactone oxidase (EC:1.3.2.3) that uses a different substrate was described in T. cruzi, which is also capable of synthesizing ascorbate (75).

The strictly specific enzymes involved in the metabolism of nucleotides and nucleic acids identified here in L. major were already reported in the literature. In L. major, a nonspecific nucleoside hydrolase (EC:3.2.2.1) capable of hydrolyzing all natural purine and pyrimidine nucleosides was described by Shi et al. (1999) (76). Subsequently other nucleoside hydrolases were identified in different trypanosomatids (77,78) and inhibitors were identified for these targets in order to design new antitrypanosomal compounds (78,79). Photolyases (EC:4.1.99.3) are flavoproteins involved in the repair of DNA damages caused by UV light. These enzymes act by reversing cyclobutane pyrimidine dimers and were described in T. cruzi (15). S1 nucleases (EC:3.1.30.1) are enzymes that hydrolyze nucleotide bonds catalyzing break of RNA and single-stranded DNA. This enzyme has been described for the first time in trypanosomatids in L. donovani and seems to have a fundamental role in the salvage of purines (80). Recently a new excreted S1 nuclease has been described in L. donovani (81).

Considering strictly specific enzymes with a higher potential for drug development, our analysis showed that the four enzyme functions: sterol 24-C-methyltransferase (EC:2.1.1.41), trypanothione synthetase (EC:6.3.1.9), pyruvate phosphate dikinase (EC:2.7.9.1) and mitochondrial RNA ligase (EC:6.5.1.3) can be taken into consideration for the development of a drug cocktail.

Unlike humans who have cholesterol in their biological membranes trypanosomatids, rather use ergosterol. Therefore, trypanosomatids present distinct enzymes in their sterol biosynthesis pathway (82). The sterol 24-C-methyltransferase (EC:2.1.1.41) catalyzes a methylation of carbon
Annexe

24, which is fundamental for the ergosterol biosynthesis. The 24-C-methyltransferase reaction is inhibited by 22,26-azasterol, which causes morphological changes and lysis to Leishmania spp. (54).

Trypanothione synthetase (EC:6.3.1.9) is fundamental to the synthesis of trypanothione, a redox metabolite of trypanosomatids involved in a number of processes like regulation of intracellular thiol redox balance, drug resistance, defense against chemical agents and oxidative stress (83). The enzymes trypanothione synthetase and trypanothione reductase are involved in the trypanothione biosynthesis and metabolism pathways, and are under studies for development of an alternative chemotherapy against trypanosomatids (84).

The pyruvate phosphate dikinase (EC:2.7.9.1) is an enzyme present in trypanosomatid glycosomes (85). It was reported that this enzyme plays a key role in maintaining the balance of ATP vs. ADP in the organelles in procyclic T. brucei (86), and in the gluconeogenesis process in amastigotes of Leishmania mexicana (87). Therefore, it is believed that the pathway including pyruvate phosphate dikinase is central in the energetic metabolism of L. major and that this enzyme deserves to be considered for leishmanicidal drugs.

The L-complex has been isolated from T. brucei and Leishmania tarentolae mitochondria. At least 16 protein components have been identified, including REL1 and REL2. The precise function of REL1 and REL2 (EC:6.5.1.3) has been described by complementation of knockout strains (88,89). In T. brucei, the conditional disruption of REL1 is lethal in vivo by affecting both U-deletion and U-insertion editing resulting in an overall decrease in RNA size (90). However, loss of REL2 has no effect on viability or on editing (91).

These enzymes act in different metabolic pathways and cellular processes: sterol biosynthesis and membrane structure (EC:2.1.1.41), glutathione metabolism and oxidative stress response (EC:6.3.1.9), energy metabolism and carbon fixation pathways in prokaryotes metabolism (EC:2.7.9.1) and RNA editing/post-transcription RNA processing (EC:6.5.1.3). Therefore, the simultaneous inhibition of these enzymes and consequently these pathways / cellular processes would likely promote an irreversible collapse of these parasites.

We verified that at the exception of EC:2.7.9.1 that has only one associated enzyme sequence, the other three EC numbers for strictly specific targets (EC:2.1.1.41, EC:6.5.1.3, EC:6.3.1.9) have more than one gene representative in the L. major genome. Gene redundancy for a same enzymatic function is a potential source of drug resistance because a given inhibitor may inhibit a given enzyme, but not its putative isoforms that could be present as well (92). Drug resistance may
involve several different mechanisms and a common solution to this problem is to increase the number of enzymes that are targeted by inhibitors at the same time.

The process of gene accumulation, which has been called pyramidation (93) is routinely used in classical breeding for plant resistance where it was first implemented (94,95). Pyramidation lowers the likelihood of virulence adaptation by a parasite because the corresponding accumulation of virulence genes becomes unsustainable in the given environmental conditions.

The idea of accumulating a number of targets to inhibit simultaneously for a better disease control is a variation of the gene-for-gene relationship that has been described by Flor (96) and can be referred to as gene-for-inhibitor concept (97). Thus, formulating drugs into a cocktail should overcome parasite resistance, which is actually a modern therapeutic trend for therapy against cancer (98). However, formulating a drug combination should also account for the dose-limiting negative side effects for normal cells and to protect the integrity of the host immune system. Thus, we sought to address this question here by identifying only enzyme targets specific to L. major since targeting these candidates would most likely minimize deleterious side effect from therapeutic combination to patients.

Drug reposition is a preferred route in drug development due to savings in time and money. Unfortunately, an examination of TDR Targets (http://tdrtargets.org/) did not show any approved drugs for the enzyme targets described here, but some inhibitors appear to be available at least for the experimental stages in Leishmania or other biological species. A search in the literature uncovered that (i) azasterols are inhibitors of sterol 24-C-methyl transferase (EC:2.1.1.41) in Leishmania (54,99); (ii) azo dye naphthalene-like compounds, similar to suramin, are inhibitors of REL (EC:6.5.1.3) (100). Two of these compounds were later tested on whole editosome, and showed strong inhibition of U-deletion RNA editing (101). GW5074, mitoxantrone, NF 023, protoporphyrin IX, and D-sphingosine were also identified as REL inhibitors, with IC50 values ranging from 1 to 3 µM (102); (iii) phosphonopyruvate (Id: DB02522 in DrugBank) has been characterized as an inhibitor (103) of pyruvate phosphate dikinase (EC:2.7.9.1) as well as flavones analogs. It has also been reported that flavones analogs act by specifically targeting the ATP binding site of the enzyme in a fungal system (104). Several phenolic compounds were also proposed for the inhibition of pyruvate phosphate dikinase in wolbachia (105); and (iv) a bis-benzyl diamine derivative as well as MOL2008 were proposed as new molecular scaffolds for drug development that proved to be active at EC50 = 12 µM on Leishmania infantum promastigotes; these compounds depleted intracellular trypanothione by inhibiting trypanothione synthetase (EC:6.3.1.9) in treated parasites confirming the on-target activity of these compounds (106).
Annexe

In the hypothesis that the susceptibility of Leishmania would be stable in vitro to a combination of the compound listed above, it would become possible to optimize their scaffold structure in order to reduce the overall cocktail toxicity to a human cell model. The engineering of this process is now facilitated by in silico modeling. A search for 3D models in PDB (http://www.rcsb.org/pdb/home/home.do) with sequences homologous to LmjF.36.2380, LmjF.36.2390, LmjF.11.1000, LmjF.23.0460, LmjF.27.1870, LmjF.36.4300, LmjF.01.0590 and LmjF.20.1730 showed that only the sequences of EC:2.7.9.1 and EC:6.3.1.9 were suitable for 3D modeling because of their identity scores larger than 39% reaching up to a maximum of 97%. However, in the case of EC:6.5.1.3 and EC:2.1.1.41, identity scores were too low (<25%) and the number of gaps too high for successful 3D modeling.

It would be interesting to conduct a phylogenetic search for the set of L. major-true positives enzymes in other species to trace the evolutionary origin of specific proteins and their properties in protozoan parasites. For instance, the sequences of trypanothione synthetase (EC: 6.3.1.9) of T. cruzi showed identities between 37% and 58% with those of L. major. These data agree with the literature, which demonstrate the expression of this enzyme and the presence of the corresponding trypanothione metabolism in T. cruzi (84). As expected, L. major, which belongs to Tritryps, had a greater phylogenetic affinity with T. cruzi and T. brucei (Euglenozoa) than with P. falciparum, which belongs to a different phylum (Apicomplexa).

Conclusions

By applying a comparative method of enzyme function for the pathogenic relationship between L. major and humans, we succeeded in greatly restricting the number of potential protein targets we believe to be suitable for drug development. The method is systematic, includes a process for determining false positives and can be automated to identify specific enzymes that strictly exist in their structural form only in the parasite. Strictly specific enzymes represent candidates for the best opportunity to engineer drugs that minimize treatment toxicity to the host. Our analysis uncovered four enzymes in L. major (sterol 24-C-methyl transferase, RNA editing ligase, pyruvate phosphate dikinase and trypanothione synthetase) whose designation as strictly specific is validated by previous published studies. We propose their use in the development of a drug cocktail to minimize the likelihood of drug resistance and the expected decrease in toxicity towards patient. Finally, we believe the same methodology can be applied in the investigation of other host-parasite relationships.
Annexe

Acknowledgements

We thank members of the Plataforma de Bioinformática Fiocruz RPT04-A/RJ for technological support and David William Provance Jr. for reading and editing the manuscript.
Annexe

References


Annexe


Annexe


Annexe


Annexe


Annexe


Annexe

http://linkinghub.elsevier.com/retrieve/pii/S1684118216300226


Annexe
SYNTHETASE PARALOG FROM LEISHMANIA DONOVANI. J Biol Chem [Internet].
2014 Apr 25;289(17):12096–108. Available from:
http://www.jbc.org/cgi/doi/10.1074/jbc.M114.554642

http://dx.plos.org/10.1371/journal.pntd.0004365

biochemical characterization, and in-vivo expression of the intracellular invertase BfrA from
Available from: http://linkinghub.elsevier.com/retrieve/pii/S0008621515001986

conserved secretory invertase is differentially expressed by promastigote developmental
forms of all species of the human pathogen, Leishmania. Mol Cell Biochem [Internet]. 2015

71. Wilkinson SR, Prathalingam SR, Taylor MC, Horn D, Kelly JM. Vitamin C biosynthesis in
trypanosomes: A role for the glycosome. Proc Natl Acad Sci [Internet]. 2005 Aug
16;102(33):11645–50. Available from:
http://www.pnas.org/cgi/doi/10.1073/pnas.0504251102

72. Adak S, Datta AK. Leishmania major encodes an unusual peroxidase that is a close homologue
http://biochemj.org/lookup/doi/10.1042/BJ20050311

Leishmania major UDP-sugar pyrophosphorylase salvages galactose for glycoconjugate
biosynthesis. Int J Parasitol [Internet]. 2015 Oct;45(12):783–90. Available from:
http://linkinghub.elsevier.com/retrieve/pii/S0020751915001861

UDP-Glucose and UDP-Galactose Using a Degron System Leads to Growth Cessation of


82. Roberts CW, McLeod R, Rice DW, Ginger M, Chance ML, Goad LJ. Fatty acid and sterol metabolism: potential antimicrobial targets in apicomplexan and trypanosomatid parasitic
Annexe


Annexe

7;7(12):e51020. Available from: http://dx.doi.org/10.1371/journal.pone.0051020


Annexe

12;10(4):e0004617. Available from: http://dx.plos.org/10.1371/journal.pntd.0004617
Annexe

Supporting Information

S1 File. Detailed explanation on the methodology of false positive search.

S2 Table. Excel sheet dashboard of homologous hit analysis for protein-protein and protein-genome comparison between L. major and H. sapiens. *This file is too big to be inserted into this manuscript.*

S3 File. Detailed explanation on the results of enzyme functional specificity.

S4 Table. List of specific proteins of L. major obtained by BLAST comparison com H. sapiens.

S5 Table. Proteins of T. cruzi, T. brucei and P. falciparum homologous to the true positives proteins of L. major.
Annexe

S1: Analysis of consistency in homologous pair detection of *L. major* vs. *H. sapiens* comparison

Because a homologous region in a protein from *L. major* and its evolutionary distant human host was expected to cover only a short DNA region (only involving a conserved protein domain) compared to the real size of its human counterpart, we considered a large human DNA region including flanking sequences of 1,000 bp on the 5’ and 3’ sides of the homologous stretch in this study. There could be several compatible genes with a distant homologous protein in the DNA region due to annotation uncertainty and the exon-intron structure of human genes. To identify the proper matches, we proceed as follow:

First, after retrieving the tBLASTn hits that were effectively associated with Ensembl proteins using an automated process (Fig 1), we identified eleven possibilities (Fig 1A), of which 7 are considered *TRUE* and 4 are considered *FALSE* in the Boolean sense. The FALSE and TRUE options can be easily diagnosed according to the decision tree shown in Fig 1B.
Fig 1. Homologous hit search algorithm for the classification of false positives of parasite specific genes in humans. A: The eleven possibilities of Ensembl protein associations (S2-E2) that one may obtain with a human genomic region homologous (tBLASTn) to a parasite protein query (S1-E2). S and E are for the beginning (start) and end of a tBLASTn homology or a human gene coordinates, respectively. The Boolean description for each association between a tBLASTn hit in a human genome region and the Ensembl proteins annotated in that region is given on the left and right sides of panel A. ‘&’ is used here in its Boolean sense, i.e., a logical AND. Human genes for Ensembl protein eventually compatible with a parasite homologous counterpart (tBLASTn) is
Annexe
modeled by thin lines. Human genes for Ensembl proteins in the same genomic region as a parasite homologous counterpart, but not compatible with it are represented by dashed lines. B: Decision tree for TRUE and FALSE associations of human genes for Ensembl proteins with a genomic region of the human genome having a homology (tBLASTn) hit for a parasite protein.

Second, *L. major* genes lack introns, which may complicate the direct comparison of its gene sequences with those of the human genome because the homologous regions, if any, could be interrupted by a splicing site in the human sequence of the homologous gene. To clarify this type of ambiguity, we checked whether a genomic hit obtained by comparing the *L. major* proteins with the DNA sequence of human chromosomes (tBLASTn), as described in the previous paragraph, also matched the homologous region detected in the BLASTp (PSI-BLAST) comparison of the same *L. major* query with the human protein sequence (subject) corresponding to that genomic hit.

Third, we mapped the homologous regions of *L. major* that were obtained through the tBLASTn and BLASTp searches to their respective human genes by comparing them to the gene structure models available from NCBI CCDS. These comparisons provided information about whether a homologous region might exactly match an exon, overlap two exons, or overlap an exon-intron boundary. This strategy also allowed us to consistently analyze the tBLASTn and BLASTp comparisons.

Thus, the compilation of the homology searches and gene models revealed the necessary information (i) to verify whether a protein domain from *L. major* was effectively present in the human genome, (ii) to map the location of the domain inside human genes, and (iii) to confirm whether the homology was a false positive or not (Fig 2). To this end, we grouped the information from all analyses in the same table and searched for a process to map the homologous regions according to gene structure. If the coordinates of the homologous regions match the positions of an exon on the same strand, one may conclude that the homology is representative of the tagged gene. In contrast, any inconsistency between the coordinates of the homologous regions and the gene
Annexe
structure or strand would lead to the conclusion that the homology under consideration is not representative of the tagged gene. This systematic process of homologous region versus a gene structure evaluation can be automated and is suitable for false positive detection. Thus, the homologies of two proteins (from *H. sapiens* and *L. major*) with the same DNA, which were considered independently, must match to be considered consistent.

![Diagram](image)

**Fig 2.** Search scheme for false positives of parasite specific genes in human. The homologous hit of *L. major* protein with the human genome is compared to CCDS annotation from Havana and the gene structure obtained by tBLASTn search with Ensembl proteins.

Regarding the automation of the processes described above, one may recognize that an Ensembl protein sequence can be mapped to its corresponding human DNA region if the homologous region between *L. major* and the human genome when (i) the initial subject coordinate (human genome) of the homologous region is larger than or equal to the initial coordinate of the human gene (corresponding to the mapped Ensembl protein) plus the initial gene coordinate of the exon in which the homology occurs and (ii) the final subject coordinate of the homologous region is smaller than or equal to the final coordinate of the gene minus the final gene coordinate of the exon
Annexe

in which the homology occurs.

To map a homologous region of an *L. major* protein on its tagged human gene, we must consider several lines of evidence. (i) Homology may be split into two homologous regions if an intron falls in the genomic region (subject) corresponding to the protein domain of the query that is involved in the homology. In this case, the two homologies of the query would be contiguous in the final coordinate of the first region and the initial coordinate of the second region. (ii) The genomic coordinate of the final position of a homologous region minus the genomic coordinate of its initial position plus one must be a multiple of three because of the codon structure of CDSs. (iii) The homologous region encodes a protein, and thus the subject coordinates must match exons.

Thus, to automate the process of homologous region mapping to human exons, (i) we compared the *L. major* proteins with human proteins (BLASTp) in Ensembl. (ii) Considering the significant match of this comparison, we recovered the coordinates of the human gene according to the chromosome sequences. (iii) To determine the gene models corresponding to human proteins that have a significant hit with *L. major* proteins, we compared (tBLASTn) the human proteins from Ensembl to their respective genomic sequence stretch by extracting them from the chromosomal sequence according to the coordinates of the Ensembl annotation. (iv) The homologies detected by the comparison under ‘iii’ gave us the coordinates for the exons (subject) relative to the gene sequence, which is outside its genomic context. (v) To translate the exon coordinates that relate to the genes (under ‘iv’) into coordinates that relate to the chromosomes, we added the exon coordinates that are related to the gene to the gene coordinates that are related to the chromosomes, as provided by Ensembl. (vi) To calculate the intron size, we subtracted the chromosomal coordinate (under ‘v’) of the beginning of an exon from the chromosomal coordinate of the end of the preceding exon for each exon of the homologous genes. (vii) To calculate the exon size, we subtracted the chromosomal coordinate (under ‘v’) for the end of an exon from the chromosomal coordinate for its beginning and added 1; we did this for all exons of the homologous genes. Of course, the size of all of the exons had to be a multiple of 3 (codons) to be considered relevant. (viii)
Annexe

The sum of the exon size divided by 3 gave us the protein size, which was inferred from the model obtained from the chromosomal sequence that could then be compared with the query protein from Ensembl for size consistency. (ix) Another consistency test was performed by determining whether the strands (‘+’ or ‘-’) from the exons deduced from the subject homologies were all ‘+’ or all ‘-’. Alternating the ‘+’ and ‘-’ strands for the exons of a same gene that was homologous to the same query protein would not make sense and allowed us to diagnose a non-significant homology. (x) We also confirmed the gene models we obtained from the tBLASTn search of the Ensembl proteins and the chromosomal DNA by comparing the coordinates of the exons that we calculated with the coordinates of the exons available in the CCDS database, as annotated by Ensembl. (xi) To map the homologous region of a *L. major* query with a human exon, we subtracted the initial and final chromosomal (human) coordinates (the subject) that corresponded to the hit based on the initial and final coordinates of each human exon (obtained in ‘v’), respectively. The exon for which both subtractions gave values equal to zero (or close to zero providing that the small protruding end was a multiple of 3 bases) was considered the one that matched the homologous region in the *L. major* query.

In addition to the AnEnPi criteria, in this study, we considered the homologies between a *L. major* protein and a human protein in the Ensembl list to be true positive when (i) the human protein was legitimate (all exons mapping on the same DNA strand), and (ii) the homologous region between both proteins (from *H. sapiens* and *L. major*) was matched in the coding frame and the coordinates with the genomic sequence corresponding to the hit of the comparison (tBLASTn) of the *L. major* proteins with the chromosomal DNA. In other words, to be considered significant, a homology between *L. major* and the human proteins had to be consistent with the gene model of the human protein, i.e., the homologous regions between the proteins and DNA had to match the exons. All of these operations were gathered together in one single Excel sheet that we used as a dashboard (see the Table in the supplementary file S2).
Annexe

Proteins domain mapping

Proteins are generally composed of one or more functional regions, commonly termed *domains*. Domains are defined as structural entities that are able to maintain their folding and activity when removed from their whole protein context. They tend to be conserved among distant organisms as a result of the selective pressures operating on them to maintain their functionality. Different combinations of these domains (cores) gave rise to a large range of protein functions in nature, and the identification of domains that occur within proteins can therefore provide insights into their function. Thus, when considering the enzymatic function of a given protein, each domain should not be considered individually but should be considered collectively as functional cores because it is their assembly that determines the final enzymatic function. Consequently, the method used to search for a core formed by several sub-domains is usually performed by detecting the location of the core within the protein by a homologous comparison with distant organisms and searching for pairs that are characterized by an identity rate of ~40% and a score rate of >120. Therefore, the sub-domain composition can be characterized by querying domain databases such as Pfam. In cases where a homology was found between *L. major* and *H. sapiens*, the homology can be immediately associated to a core because the organisms are distant from each other. When drug development is considered for specific enzymes that do not have homologies in *H. sapiens*, the core responsible for the enzymatic function must be analyzed in other organisms, such as plants. To map the domains in the conserved regions of *H. sapiens* and *L. major* homologous proteins, we used the Pfam database (http://pfam.xfam.org/search) and the Conserved Domains Database (CDD) (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). CDD is a protein annotation resource that consists of a collection of well annotated multiple sequence alignment models for ancient domains and full-length proteins. The Pfam and CDD databases also provide canonical tertiary structures (3D) for mapped domains that are referenced to the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do). In an effort to identify the possible *functional specificities* in the homologous domains of *L. major* and *H. sapiens*, we characterized the
Annexe

differences between the secondary structures (2D) of *L. major* and *H. sapiens* using PRALINE (http://www.ibi.vu.nl/programs/pralinewww/).
Annexe

S3: Homologous proteins classified as functional specific enzymes

We recovered 80 coding sequences (CDS) that were annotated as consensus coding sequences (CCDSs) by Ensembl and reported to be effectively translated into proteins from the genomic coordinates within the human DNA regions corresponding to the 55 tBLASTn hits. By comparing the coordinates of the gene structures (exons + introns) reported by Ensembl for these CCDSs with the coordinates of the homologies given by the tBLASTn search of their protein sequence with the human genome (putative exons), we identified many inconsistencies in the Ensembl CCDS annotations. By analyzing the 55 hits, we identified 41 non-consistent protein cases where putative exons were (i) associated with the same protein but appeared on both gene strands, which is impossible given the translation process, or (ii) were simply deprived of a CCDS association, making any further inference impossible (Table S1).

Table 1. Protein clusters and the paralogy or analogy relationship between each cluster.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>L. major protein identifier</th>
<th>EC number</th>
<th>Gene location</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Chromosome number</td>
<td>Start</td>
</tr>
<tr>
<td>1</td>
<td>LmjF:24.2030</td>
<td>1.1.1.100***</td>
<td>24</td>
<td>752837</td>
</tr>
<tr>
<td></td>
<td>LmjF:27.2440</td>
<td>1.1.1.100***</td>
<td>27</td>
<td>1079832</td>
</tr>
<tr>
<td>2</td>
<td>LmjF:36.2380</td>
<td>2.1.1.41**</td>
<td>36</td>
<td>961263</td>
</tr>
<tr>
<td></td>
<td>LmjF:36.2390</td>
<td>2.1.1.41**</td>
<td>36</td>
<td>965198</td>
</tr>
<tr>
<td>3</td>
<td>LmjF:31.2300</td>
<td>3.1.30.1**</td>
<td>31</td>
<td>1138947</td>
</tr>
<tr>
<td></td>
<td>LmjF:31.2310</td>
<td>3.1.30.1**</td>
<td>31</td>
<td>1141911</td>
</tr>
<tr>
<td>4</td>
<td>LmjF:28.2100</td>
<td>3.1.4.46****</td>
<td>28</td>
<td>819522</td>
</tr>
<tr>
<td></td>
<td>LmjF:36.5960</td>
<td>3.1.4.46****</td>
<td>36</td>
<td>2296398</td>
</tr>
<tr>
<td>5</td>
<td>LmjF:04.0310</td>
<td>3.2.1.26**</td>
<td>4</td>
<td>103123</td>
</tr>
<tr>
<td></td>
<td>LmjF:04.0320</td>
<td>3.2.1.26**</td>
<td>4</td>
<td>106566</td>
</tr>
<tr>
<td></td>
<td>LmjF:23.0870</td>
<td>3.2.1.26**</td>
<td>23</td>
<td>430918</td>
</tr>
<tr>
<td></td>
<td>LmjF:23.0880</td>
<td>3.2.1.26**</td>
<td>23</td>
<td>435934</td>
</tr>
<tr>
<td></td>
<td>LmjF:35.0640</td>
<td>3.2.1.26**</td>
<td>35</td>
<td>302145</td>
</tr>
<tr>
<td>6</td>
<td>LmjF:14.0180</td>
<td>3.4.17.19**</td>
<td>14</td>
<td>46644</td>
</tr>
<tr>
<td></td>
<td>LmjF:33.2540</td>
<td>3.4.17.19**</td>
<td>33</td>
<td>1138885</td>
</tr>
<tr>
<td>7</td>
<td>LmjF:10.0460</td>
<td>3.4.24.36**</td>
<td>10</td>
<td>213345</td>
</tr>
<tr>
<td></td>
<td>LmjF:10.0465</td>
<td>3.4.24.36**</td>
<td>10</td>
<td>216396</td>
</tr>
<tr>
<td></td>
<td>LmjF:10.0470</td>
<td>3.4.24.36**</td>
<td>10</td>
<td>219443</td>
</tr>
<tr>
<td></td>
<td>LmjF:10.0480</td>
<td>3.4.24.36**</td>
<td>10</td>
<td>223401</td>
</tr>
<tr>
<td></td>
<td>LmjF:28.0570</td>
<td>3.4.24.36**</td>
<td>28</td>
<td>201489</td>
</tr>
<tr>
<td></td>
<td>LmjF:31.2000</td>
<td>3.4.24.36**</td>
<td>31</td>
<td>979311</td>
</tr>
<tr>
<td>8</td>
<td>LmjF:18.0560</td>
<td>3.6.3.6****</td>
<td>18</td>
<td>228469</td>
</tr>
<tr>
<td></td>
<td>LmjF:18.1510</td>
<td>3.6.3.6****</td>
<td>18</td>
<td>679523</td>
</tr>
</tbody>
</table>
### Table

<table>
<thead>
<tr>
<th>Gene</th>
<th>Enzyme ID</th>
<th>EC Numbers</th>
<th>Gene Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>LmjF.18.1520</td>
<td>3.6.3.6****</td>
<td>18</td>
<td>684689</td>
</tr>
<tr>
<td>LmjF.09.0360</td>
<td>4.1.99.3**</td>
<td>9</td>
<td>145424</td>
</tr>
<tr>
<td>LmjF.33.0470</td>
<td>4.1.99.3**</td>
<td>33</td>
<td>207188</td>
</tr>
<tr>
<td>LmjF.19.1020</td>
<td>4.2.1.70****</td>
<td>19</td>
<td>420290</td>
</tr>
<tr>
<td>LmjF.26.0420</td>
<td>4.2.1.70****</td>
<td>26</td>
<td>114341</td>
</tr>
<tr>
<td>LmjF.36.1660</td>
<td>4.2.1.70****</td>
<td>36</td>
<td>652829</td>
</tr>
<tr>
<td>LmjF.30.1550</td>
<td>4.2.1.70****</td>
<td>30</td>
<td>548003</td>
</tr>
<tr>
<td>LmjF.19.0985</td>
<td>6.2.1.12**</td>
<td>19</td>
<td>398614</td>
</tr>
<tr>
<td>LmjF.19.0995</td>
<td>6.2.1.12**</td>
<td>19</td>
<td>403723</td>
</tr>
<tr>
<td>LmjF.19.1005</td>
<td>6.2.1.12**</td>
<td>19</td>
<td>410749</td>
</tr>
<tr>
<td>LmjF.23.0460</td>
<td>6.3.1.9**</td>
<td>23</td>
<td>174520</td>
</tr>
<tr>
<td>LmjF.27.1870</td>
<td>6.3.1.9**</td>
<td>27</td>
<td>806164</td>
</tr>
<tr>
<td>LmjF.36.4300</td>
<td>6.3.1.9**</td>
<td>36</td>
<td>1657674</td>
</tr>
<tr>
<td>LmjF.01.0590</td>
<td>6.5.1.3**</td>
<td>01</td>
<td>163208</td>
</tr>
<tr>
<td>LmjF.20.1730</td>
<td>6.5.1.3**</td>
<td>20</td>
<td>715134</td>
</tr>
</tbody>
</table>

* No hit with the enzymes in the same cluster.

** Specific enzyme of *L. major* in a cluster.

*** Homologous enzyme of *L. major* in a cluster.

**** Enzymes that are common to both specific and unspecific set of *L. major*.

The comparison of the gene structure obtained (i) by homologous comparison between the Ensembl proteins and human DNA and (ii) from the putative exons reported in the CCDS annotations allowed us to diagnose clear CCDS inconsistencies and, consequently, to identify those that might represent true homologies between *L. major* and humans. The 15 unspecific *L. major* proteins could be classified into three categories according to the EC numbers of their homologous human counterparts (Table S2). Among the 15 *L. major* enzymes, 14 demonstrated homology to the proteins from Ensembl and one (LmjF.23.0270) to the human genome sequence itself. The 14 *L. major* enzymes were homologous to 17 human proteins from the Ensembl set. Among 17 human homologies, nine were annotated with different EC numbers compared to their *L. major* counterpart, which mean that they were associated to different enzymatic reactions in *L. major* and humans and, therefore, could be considered specific from a functional standpoint (if their enzymatic annotation is correct). Among the eight remaining *H. sapiens* homologous pairs, three were associated to enzyme assigned with incomplete EC number, which explain why they were associated to different enzymatic activities during the process of AnEnPi clustering. For the five...
Annexe

remaining cases, the *L. major* sequences were annotated as enzymes but their human homologous pairs were not, which may explain why they passed through the AnEnPi filtering system as potentially specific (see Table 2). It is worth noting here that in all 17 cases, the homologous region between the *L. major* protein and the human genome sequences precisely corresponded to a unique exon. These conserved regions corresponding to the protein domains were never interrupted by an intron in the human genome.

Table 2. EC number annotations of human homologous proteins for *L. major* enzymes classified as functionally specific by AnEnPi.

<table>
<thead>
<tr>
<th>Types</th>
<th>EC number annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different EC numbers</td>
<td>LmjF.36.5960* - EC:3.1.4.46 (Glycerophosphodiester phosphodiesterase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:51573* - EC:3.1.4.44 (Glycerophosphoinositol glycerophosphodiesterase)</td>
</tr>
<tr>
<td></td>
<td>LmjF.35.4250 - EC:2.1.1.64 (tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:51805 - EC:2.1.1.114 (polyprenyldihydroxybenzoate methyltransferase)</td>
</tr>
<tr>
<td></td>
<td>LmjF.28.2910 - EC:1.4.1.4 (glutamate dehydrogenase (NADP+))</td>
</tr>
<tr>
<td></td>
<td>Hsa:2747 - EC:1.4.1.3 (glutamate dehydrogenase [NAD(P)+])</td>
</tr>
<tr>
<td></td>
<td>LmjF.18.1510 - EC:3.6.3.6 (H+-exporting ATPase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:478 - EC:3.6.3.9 (Na+/K+-exchanging ATPase)</td>
</tr>
<tr>
<td></td>
<td>LmjF.18.1520 - EC:3.6.3.6 (H+-exporting ATPase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:478 - EC:3.6.3.9 (Na+/K+-exchanging ATPase)</td>
</tr>
<tr>
<td></td>
<td>LmjF.33.0680 - EC:1.3.1.71 (Delta24(241)-sterol reductase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:7108 - EC:1.3.1.70 (Delta14-sterol reductase)</td>
</tr>
<tr>
<td></td>
<td>LmjF.05.0350 - EC:1.8.1.12 (Trypanothione reductase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:7296 - EC:1.8.1.9 (Thioredoxin reductase (NADPH))</td>
</tr>
<tr>
<td></td>
<td>LmjF.09.0770 - EC:3.4.21.83 (Oligopeptidase B)</td>
</tr>
<tr>
<td></td>
<td>Hsa:5550 - EC:3.4.21.26 (Prolyl oligopeptidase)</td>
</tr>
<tr>
<td></td>
<td>LmjF.30.1550 - EC:4.2.1.70 (pseudouridylate synthase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:83480 - EC:5.4.99.45 (tRNA pseudouridine (38/39) synthase)</td>
</tr>
<tr>
<td>Incomplete EC number</td>
<td>LmjF.26.2280 - EC:3.5.5.1 (nitrilase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:4817 - EC:3.5.-,- (nitrilase 1)</td>
</tr>
<tr>
<td></td>
<td>LmjF.24.2030 - EC:1.1.1.100 (3-oxoacyl-[acyl-carrier-protein] reductase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:84869 - EC:1.1.1.- (oxidoreductase)</td>
</tr>
<tr>
<td></td>
<td>LmjF.27.2440 - EC:1.1.1.100 (3-oxoacyl-[acyl-carrier-protein] reductase)</td>
</tr>
<tr>
<td></td>
<td>Hsa:84869 - EC:1.1.1.- (oxidoreductase)</td>
</tr>
<tr>
<td>EC number not assigned</td>
<td>LmjF.14.0350 - EC:4.2.3.1 (threonine synthase)</td>
</tr>
<tr>
<td></td>
<td>Hsa: N/A.</td>
</tr>
<tr>
<td></td>
<td>LmjF.30.1550 - EC:4.2.1.70 (pseudouridylate synthase)</td>
</tr>
<tr>
<td></td>
<td>Hsa: N/A.</td>
</tr>
<tr>
<td></td>
<td>LmjF.30.0180 - EC:1.1.1.60 (2-hydroxy-3-oxopropionate reductase)</td>
</tr>
</tbody>
</table>
Annexe

| Hsa: N/A. |
|---|---|
| LmjF.18.1510 - EC: 3.6.3.6 (H+-exporting ATPase) |
| Hsa: N/A. |
| LmjF.18.1520 - EC: 3.6.3.6 (H+-exporting ATPase) |
| Hsa: N/A. |

*Accession numbers from KEGG.*

For seven hits (hits of Table 3 plus LmjF.26.2280) from the 55 obtained through tBLASTn comparison of the *L. major* enzymes (homologous set) with the human genome, no Ensembl data for the human protein was available; three (LmjF.18.1510, LmjF.18.1520 and LmjF.27.2440) of five proteins matched a protein from other *H. sapiens* genomic coordinates, suggesting that they could be pseudogenes that arose from a previous sequence duplication event.

Two sequences that did not match any Ensembl protein were LmjF.23.0270 and LmjF.26.2280. The protein sequence LmjF.26.2280 matched a pseudogene since a ‘*’ could be found in the sequence of the human homologous region in the tBLASTn alignment (subject). This human pseudogene corresponded to a homology with LmjF.26.2280 (EC: 3.5.5.1), which has been described to act as a nitrilase. For the other six *L. major* hits with the human genome, we could retrieve a protein function for the human conserved region using the sequence of the subject homologous region (human) in the BLAST alignment as a query to search (BLASTp) the *nr* section of GenBank (Table 3).

Table 3. *H. sapiens* proteins retrieved from *nr* using the DNA stretch, of the subject corresponding to the homologous region between *L. major* proteins and the human genome, as a query.

<table>
<thead>
<tr>
<th><em>L. major</em> identifier</th>
<th>Chromosome number</th>
<th>Hit genomic coordinate</th>
<th>Human homologous proteins in <em>nr</em></th>
<th>Function</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>LmjF.18.1510 and LmjF.18.1520</td>
<td>8</td>
<td>19098173-19098421</td>
<td>PMCA2a, partial</td>
<td>NP_064524</td>
<td></td>
</tr>
<tr>
<td>LmjF.18.1510 and LmjF.18.1520</td>
<td>18</td>
<td>19998852-19999097</td>
<td>hCG2039601</td>
<td>EAX0113</td>
<td></td>
</tr>
<tr>
<td>LmjF.23.0270</td>
<td>6</td>
<td>99622932-</td>
<td>3-hydroxybutyrate</td>
<td>AAH01953</td>
<td></td>
</tr>
</tbody>
</table>
However, we could not confirm the homology of the human proteins identified by tBLASTn with the human genomic DNA, including the homologous region with the \textit{L. major} protein plus or minus 1,000 bp. If the six human proteins retrieved from \textit{nr} were actually homologous to the considered human DNA stretches, one would expect that they should produce homologous hits. Because this was not the case, we concluded that the seven homologous hits of \textit{L. major} in the human genome did not match any significant protein that could have escaped the Ensembl annotation process and should likely considered as paralogous genes that evolved into pseudogenes.

**Protein domain mapping**

To better characterize the functional specificities, we searched the catalytic cores of 12 of the 14 enzymes of \textit{L. major} described in Table 2. We searched for their distant homologies with protein sequences in the \textit{nr} database using identity and score rates of \(~40\%\) and \(\geq120\), respectively. We mapped their domain composition using Pfam and CDD. Because CDD imports external source databases (Pfam, SMART, COG, PRK, and TIGRFAMs), we identified the same 16 domains from both Pfam and CDD in five \textit{L. major} functionally specific enzymes (LmjF.05.0350, LmjF.09.0770, LmjF.18.15.10, LmjF.28.2910, LmjF.30.0180). For the other seven functionally specific enzymes (LmjF.14.0350, LmjF.24.2030, LmjF.26.2280, LmjF.30.1550, LmjF.33.0680, LmjF.35.4250, LmjF.36.5960), the domains had only approximate matches between Pfam and CDD, despite their similar functions.

The alignments of the predicted secondary structures for the 14 \textit{L. major} proteins with their homologous human counterparts showed in the conserved regions several minute differences in the location of the \(\alpha\)-helices and \(\beta\)-sheets between pairs characterized by an incomplete EC number (EC:3.5.-.-.-). We found that the predicted \(\alpha\)-helix and \(\beta\)-sheet distribution was very similar for both \textit{H. sapiens} and \textit{L. major}. In the alignments where sequences were annotated with different
Annexe

substrates (see Table 2) Similarly, in homologous pairs with two different EC numbers, the distribution pattern of the predicted secondary structures was also very similar.

For the homologous pairs (LmjF.14.0350 - EC:4.2.3.1, LmjF.30.1550 - EC:4.2.1.70, LmjF.30.0180 - EC:1.1.1.60, LmjF.18.1510 - EC:3.6.3.6, LmjF.18.1520 - EC:3.6.3.6) lacking an annotation of enzymatic activities and for which we could not recover any function for the *H. sapiens* counterpart, the *L. major* protein was composed of more \( \beta \)-sheets than \( \alpha \)-helices, which did not occur in *H. sapiens*, where the ratio between \( \beta \)-sheets and \( \alpha \)-helices was approximately the same.

The subdomains that were mapped by comparing the *L. major* and *H. sapiens* domains obtained by BLAST homology with those of CDD and Pfam showed a better consistency for Pfam (the Pfam subdomains were ordered without overlapping), which was expected because CDD integrates the data from several independent databases (including Pfam). Due to the similarities in the secondary structure composition and order in the 2D alignments, we were not able to determine whether there was any obvious functional specificity in the homologous protein pairs from *L. major* and *H. sapiens*. In addition, the absence of homologous pairs in the human sequences through the HHM search prevented a further comparison using 3D alignment.

**Metabolic pathway significance**

When possible, we associated the EC numbers of each protein corresponding to *L. major* and *H. sapiens* homologous pairs to their respective metabolic maps from the KEGG Pathway database and analyzed them according to whether they (i) had the same substrates, (ii) had different EC numbers, (iii) had incomplete EC numbers, and (iv) lacked an EC number, as detailed in Table 2.

Considering the case of homologous pairs for which the EC numbers differed in the last digit, we found that both EC:1.3.1.71 (*L. major*) and EC:1.3.1.70 (*H. sapiens*) were associated with essential pathways in these organisms, the steroid biosynthesis pathway, but EC:1.3.1.71 catalyzes the transformation of ergosterol + NADP\(^+\) into ergosta-5,7,22,24(28)-tetraen-3beta-ol + NADPH +
Annexe

H+ (map from KEGG - lma00100), while 1.3.1.70 catalyzes the transformation of 4,4-dimethyl-cholesta,8,14,24-trienol into 14-dimethyl-lanostrol in the cholesterol pathway (map from KEGG - hsa00100). Another similar case concerns EC:1.8.1.12 (L. major) and EC:1.8.1.9 (H. sapiens), where the first enzyme (trypanothione reductase) catalyzes the essential transformation of trypanothione + NADP+ into trypanothione disulfide + NADPH + H+ (glutathione metabolism: map lma00480) and the second (thioredoxin reductase) catalyzes the essential transformation of thioredoxin + NADP+ into thioredoxin disulfide + NADPH + H+ in the pyrimidine (map hsa00240) and selenocompound metabolic pathways (map lma00450).

EC:3.1.4.46 (L. major) differs from EC:3.1.4.44 (H. sapiens) in the substrates that both enzymes process. The enzyme from L. major is responsible for the essential catalysis of glycerophosphodiester + H2O into an alcohol + sn-glycerol 3-phosphate in glycerophospholipid metabolism (map lma00564), while the enzyme from H. sapiens is responsible for the essential catalysis of 1-(sn-glycero-3-phospho)-1d-myoinositol + H2O into myo-inositol + sn-glycerol 3-phosphate, which are involved in glycerol and lipid metabolism.

Concerning the homologous enzymes whose EC numbers only differ in the digit in the last position, let us note that EC:3.6.3.6 (L. major) generates an electrochemical gradient across the plasma membrane due to a difference in proton concentrations, promoting the essential reaction of ATP + H2O + H+ in into ADP + phosphate + H+ out, which is involved in the oxidative phosphorylation pathway (map lma00190). In contrast, the human counterpart, EC:3.6.3.9, is a plasma membrane enzyme that generates an ATP-dependent electrical potential by catalyzing the reaction of ATP + H2O + Na+ in + K+ out into ADP + phosphate + Na+ out + K+ in in the following pathways: (i) cGMP-PKG (map hsa04022) and cAMP (map hsa04024) signaling; (ii) cardiac muscle contraction (map hsa04260), adrenergic signaling in cardiomyocytes (map hsa04261), insulin secretion (map hsa04911), thyroid hormone synthesis (map hsa04918), thyroid hormone signaling (map hsa04919), aldosterone-regulated sodium reabsorption (map hsa04960), endocrine and other factor-regulated calcium reabsorption (map hsa04961), proximal tubule bicarbonate
Annexe

reclamation (map has04964); (iii) salivary (map has04970), gastric acid (map has04971), and pancreatic (map has04972) secretions; (iv) carbohydrates (map has04973) and proteins (map has04974) digestion and absorption; (v) bile secretion (map has04976); and (vi) mineral absorption (map has04978).

EC:1.1.1.100 (L. major) exhibits a marked preference for acyl-carrier protein derivatives over CoA derivatives as substrates, catalyzing the essential reaction (3R)-3-hydroxyacyl-[acyl-carrier protein] + NADP⁺ into a 3-oxoacyl-[acyl-carrier protein] + NADPH + H⁺, which is present in fatty acid biosynthesis (map lma00061), biotin metabolism (map lma00780), and biosynthesis of unsaturated fatty acids (map lma01040) metabolism. In H. sapiens, the homologous counterpart (EC:1.1.1.-) is an oxidoreductase that acts on organic groups of OH donors with NAD⁺ or NADP⁺ as acceptors and is involved in butanoate (map00650), pentose phosphate (map00030), pyrimidine (map00240), steroid hormone (map00140), vitamin B6 (map00750), pentose and glucuronate interconversions (map00040).

Metabolic maps are not currently available in the literature for the homologous pair of EC:3.4.21.83 (L. major) and EC:3.4.21.26 (H. sapiens), although we know that they hydrolyze specific amino acid bonds in oligopeptides. The enzyme EC:3.4.21.83 will hydrolyze a reaction after an arginine and a lysine, while the EC:3.4.21.26 enzyme is an endopeptidase that will break the peptide bonds after a proline or an alanine.

EC:2.1.1.64 (L. major) catalyzes the essential reaction of s-adenosyl-L-methionine + 3-demethylubiquinol-n into s-adenosyl-L-homocysteine + ubiquinol-n in the ubiquinone and other terpenoid-quinone biosynthesis pathways (map lma00130). The H. sapiens counterpart is EC:2.1.1.114 that is also involved in ubiquinone biosynthesis (map hsa00130) and catalyzes the essential reaction of s-adenosyl-L-methionine + 3,4-dihydroxy-5-all-trans-polyprenylbenzoate into s-adenosyl-L-homocysteine + 3-methoxy-4-hydroxy-5-all-trans-polyprenylbenzoate. However, the H. sapiens enzyme has a second EC number (EC:2.1.1.64), which means that it is also able to perform the same function as that of L. major (false positive).
Annexe

EC:1.4.1.4 (L. major) catalyzes the reaction of L-glutamate + H₂O + NADP+ into 2-oxoglutarate + NH₃ + NADPH + H⁺ in alanine, aspartate and glutamate metabolism (map lma00250), arginine and proline metabolism (map lma00330), and nitrogen metabolism (map lma00910). EC:1.4.1.3 (H. sapiens) is responsible for the same reaction as L. major and is involved in the same metabolic pathways. Therefore, one must consider that this homologous pair is a false positive according to the concept of functional specificity.

In the homologous pairs with different EC numbers, the activity of EC:4.2.1.70 (L. major) is associated with a pseudouridylate synthase, which catalyzes the essential reaction of uracil with D-ribose 5-phosphate to produce pseudo-uridine 5’-phosphate + H₂O in pyrimidine metabolism (map lma00240). Unfortunately, the map of the metabolic pathway for EC:5.4.99.45, the homologous human enzyme for EC:4.2.1.70, was not available, but this EC number is described as being involved in the formation of pseudo-uridine at the anticodon stem and loop of transfer RNAs. In addition, pseudo-uridine is an isomer of uridine 5-(beta-D-ribofuranosyl) uracil, and is the most abundant modified nucleoside in all cellular RNAs. The truA-like proteins also exhibit a conserved sequence with a strictly conserved aspartic acid, and are likely involved in catalysis.

We found several cases where the EC number of the human homologous protein was missing, which means that an enzymatic function has yet to be described for the human counterpart (i) EC:3.6.3.6 (L. major) corresponds to the H⁺-exporting ATPase. This enzyme participates in the oxidative phosphorylation pathway and promotes the reaction of ATP + H₂O + H⁺ into ADP + phosphate + H⁺ (map lma00190). (ii) EC: 4.2.3.1 (threonine synthase of L. major) catalyzes the essential reaction of o-phospho-L-homoserine + H₂O into L-threonine + phosphate in glycine, serine and threonine metabolism (map lma00260), vitamin B6 metabolism (lma00750), the biosynthesis of secondary metabolites (map01110), and the biosynthesis of amino acids (01230). (iii) EC: 4.2.1.70 (pseudouridylate synthase of L. major) catalyzes the essential reaction of uracil + D-ribose 5-phosphate into pseudo-uridine 5’-phosphate + H₂O in pyrimidine metabolism (map lma00240). (iv) EC: 1.1.1.60 (2-hydroxy-3-oxopropionate reductase of L. major) acts on the
Annexe

organic groups of OH donors using NAD⁺ or NADP⁺ as acceptors and is involved in the reaction of D-glycerate + NAD(P)⁺ into 2-hydroxy-3-oxopropanoate + NAD(P)H + H⁺ in the glyoxylate and dicarboxylate metabolic pathway (map lma00630).

When the human protein of the homologous pair had an incomplete EC number, the pathway maps could only be retrieved for *L. major*. Thus, we could only find maps for EC: 3.5.5.1, which is located in tryptophan metabolism (map lma00380), cyanoamino acid metabolism (map lma00460), and nitrogen metabolism (map lma00910). This enzyme acts on a wide range of aromatic nitriles, including (indol-3-yl) acetonitrile, some aliphatic nitriles, and the corresponding acid amides.
Annexe

S4: List of specific proteins of *L. major* obtained by BLAST comparison com *H. sapiens*.

<table>
<thead>
<tr>
<th><em>L. major</em> protein</th>
<th>EC number</th>
</tr>
</thead>
<tbody>
<tr>
<td>LmjF.07.0260</td>
<td>1.1.1.3</td>
</tr>
<tr>
<td>LmjF.17.1360</td>
<td>1.1.3.8</td>
</tr>
<tr>
<td>LmjF.34.0070</td>
<td>1.11.1.11</td>
</tr>
<tr>
<td>LmjF.26.2610</td>
<td>1.14.13.1</td>
</tr>
<tr>
<td>LmjF.16.0530</td>
<td>1.3.3.1</td>
</tr>
<tr>
<td>LmjF.30.0610</td>
<td>1.7.1.1</td>
</tr>
<tr>
<td>LmjF.31.0010</td>
<td>2.1.1.14</td>
</tr>
<tr>
<td>LmjF.36.2380</td>
<td>2.1.1.41</td>
</tr>
<tr>
<td>LmjF.36.2390</td>
<td>2.1.1.41</td>
</tr>
<tr>
<td>LmjF.09.1040</td>
<td>2.3.1.158</td>
</tr>
<tr>
<td>LmjF.36.3590</td>
<td>2.5.1.47</td>
</tr>
<tr>
<td>LmjF.30.3080</td>
<td>2.7.1.39</td>
</tr>
<tr>
<td>LmjF.17.1160</td>
<td>2.7.7.64</td>
</tr>
<tr>
<td>LmjF.11.1000</td>
<td>2.7.9.1</td>
</tr>
<tr>
<td>LmjF.36.4640</td>
<td>3.1.1.73</td>
</tr>
<tr>
<td>LmjF.31.2300</td>
<td>3.1.30.1</td>
</tr>
<tr>
<td>LmjF.31.2310</td>
<td>3.1.30.1</td>
</tr>
<tr>
<td>LmjF.04.0310</td>
<td>3.2.1.26</td>
</tr>
<tr>
<td>LmjF.04.0320</td>
<td>3.2.1.26</td>
</tr>
<tr>
<td>LmjF.23.0870</td>
<td>3.2.1.26</td>
</tr>
<tr>
<td>LmjF.23.0880</td>
<td>3.2.1.26</td>
</tr>
<tr>
<td>LmjF.35.0640</td>
<td>3.2.1.26</td>
</tr>
<tr>
<td>LmjF.29.2800</td>
<td>3.2.2.1</td>
</tr>
<tr>
<td>LmjF.17.0140</td>
<td>3.4.11.14</td>
</tr>
<tr>
<td>LmjF.14.0180</td>
<td>3.4.17.19</td>
</tr>
<tr>
<td>LmjF.33.2540</td>
<td>3.4.17.19</td>
</tr>
<tr>
<td>LmjF.10.0460</td>
<td>3.4.24.36</td>
</tr>
<tr>
<td>LmjF.10.0465</td>
<td>3.4.24.36</td>
</tr>
<tr>
<td>LmjF.10.0470</td>
<td>3.4.24.36</td>
</tr>
<tr>
<td>LmjF.10.0480</td>
<td>3.4.24.36</td>
</tr>
<tr>
<td>LmjF.28.0570</td>
<td>3.4.24.36</td>
</tr>
<tr>
<td>LmjF.31.2000</td>
<td>3.4.24.36</td>
</tr>
<tr>
<td>LmjF.07.0270</td>
<td>3.5.1.16</td>
</tr>
<tr>
<td>LmjF.34.3250</td>
<td>4.1.1.74</td>
</tr>
<tr>
<td>LmjF.09.0360</td>
<td>4.1.99.3</td>
</tr>
<tr>
<td>LmjF.33.0470</td>
<td>4.1.99.3</td>
</tr>
<tr>
<td>LmjF.14.0460</td>
<td>4.4.1.8</td>
</tr>
<tr>
<td>LmjF.18.0200</td>
<td>5.4.99.9</td>
</tr>
<tr>
<td>LmjF.19.0985</td>
<td>6.2.1.12</td>
</tr>
<tr>
<td>LmjF.19.0995</td>
<td>6.2.1.12</td>
</tr>
<tr>
<td>LmjF.19.1005</td>
<td>6.2.1.12</td>
</tr>
<tr>
<td>LmjF.26.0830</td>
<td>6.3.1.1</td>
</tr>
<tr>
<td>LmjF.23.0460</td>
<td>6.3.1.9</td>
</tr>
</tbody>
</table>
The enzymatic activity of EC:1.1.1.3 was annotated as homoserine dehydrogenase-like of L. major, LmjF.07.0260, catalyzes the reaction of L-homoserine + NAD(P)+ in L-aspartate 4-semialdehyde + NAD(P)H + H+ and could be observed on the maps related to Glycine, serine and threonine metabolism and Cysteine and methionine metabolism (lma00260 e lma00270) from KEGG.

The EC:1.1.3.8 correspond to putative L-gulonolactone oxidase of L. major, LmjF.17.1360. The overall reaction is L-gulono-1,4-lactone + O2 = L-ascorbate + H2O2 and it was performed in a two steps reaction: (a) L-gulono-1,4-lactone + O2 = L-xylo-hex-2-ulono-1,4-lactone + H2O2 and (b) L-xylo-hex-2-ulono-1,4-lactone = L-ascorbate (spontaneous). Both reactions belong to Ascorbate and aldarate metabolism (lma00053) of KEGG.

The enzymatic function EC:1.11.1.11 is present in Ascorbate and aldarate metabolism and Glutathione metabolism (lma00053 e lma00480). In the first, this function act catalyzing the reaction Ascorbate + Hydrogen peroxide <=> Dehydroascorbate + 2 H2O, while in the Glutathione metabolism, besides generating dehydroascorbate is also responsible for catalyzing the first and second steps in a two-step reaction. In the first step, the reaction catalyzes 2 Ascorbate + Hydrogen peroxide <=> 2 Monodehydroascorbate + 2 H2O and in the second catalyzes Monodehydroascorbate <=> Dehydroascorbate + Ascorbate.

The enzymatic function EC:1.14.13.1 is responsible for the reaction salicylate + NADH + 2 H + + O2 = catechol + NAD + + H2O + CO2 participating in the Dioxin degradation, Polycyclic aromatic hydrocarbon degradation and Naphthalene degradation (ec00621, ec00624 and ec00626). Although for LmjF.26.2610 the function is described in databases as TritrypDB, UniProtKB and GeneDB in the KEGG database, this protein does not have a number of EC assigned.
Annexe

The function EC:1.3.3.1 concerning specific enzyme dihydroorotate dehydrogenase from *L. major*, LmjF.16.0530 catalyzes the reaction (S) -Dihydroorotate $\Rightarrow$ Orotate + + Fumarate Succinate that is present in pyrimidine metabolism (lma00240).

The EC:1.7.1.1 is present in Nitrogen metabolism (ec00910) catalyzing the reaction Nitrite + NAD + + H2O $\Rightarrow$ Nitrate + NADH + H +, however, the enzyme, putative nitrate reductase LmjF.30.0610, does not have a number of EC allocated in the KEGG database, although it has already given function in other databases such as TritrypDB, UniProtKB and GeneDB.

The EC:2.1.1.14 catalyzes the Methyltetrahydropteroyltri-5-L-glutamate + L-Homocysteine $\Rightarrow$ Tetrahydropteroyltri-L-glutamate L-Methionine + reaction of *L. major*, LmjF.31.0010, through the 5-methyltetrahydropteroyltrim glutamate --- homocysteine S-methyltransferase enzyme that participates in Cysteine and methionine metabolism and Selenocompound metabolism (lma00270 and lma00450).

The EC:2.1.1.41 determines the function of two enzymes of *L. major* called sterol 24-C-methyltransferase, LmjF.36.2380 and LmjF.36.2390. This function is responsible for catalyzing the S-Adenosyl-L-methionine + Zymosterol $\Rightarrow$ S-Adenosyl-L-homocysteine + Fecosterol reaction engaged in Steroid biosynthesis (lma00100).

The EC:2.3.1.158 catalyzes the reaction Phospholipid + 1,2-diacyl-sn-glycerol $\Rightarrow$ Lysophospholipid + triacylglycerol by specific enzyme of *L. major*, LmjF.09.1040, putative phospholipid: diacylglycerol acyltransferase that is present in Glycerolipid metabolism (lma00561).

The EC:2.5.1.47 which has cysteine synthase function in *L. major*, LmjF.36.3590, catalyzes the reaction O-Acetyl-L-serine + Hydrogen sulfide $\Rightarrow$ Acetate + L-Cysteine which is involved in methionine and cysteine metabolism and Sulfur metabolism (lma00270 and lma00920).

The enzymatic function EC:2.7.1.39 is present in the Glycine, serine and threonine metabolism (lma00260) catalyzes the reaction ATP + L-homoserine $\Rightarrow$ ADP + O-Phospho-L-homoserine in *L. major* and is performed by the enzyme, LmjF .30.3080, putative kinase homoserine.
The EC:2.7.7.64 refers to UDP-sugar pyrophosphorylase enzyme of LmjF.17.1160 specific enzyme that catalyzes the reaction UTP + alpha-D-Aldose 1-Diphosphate ⇌ phosphate + UDP-sugar which is present in galactose metabolism, Ascorbate and aldarate metabolism, Amino sugar and nucleotide sugar metabolism and Pentose and glucuronate interconversions (lma00052, lma00053, lma00520, lma00040).

Enzymatic activity EC:2.7.9.1 for LmjF.11.1000 enzyme called pyruvate phosphate dikinase promotes the reaction ATP + Pyruvate + orthophosphate ⇌ AMP + phosphoenolpyruvate + Diphosphate in Pyruvate metabolism (lma00620).

The enzymatic function EC:3.1.1.73 is responsible for catalyzing the reaction Feruloyl-polysaccharide + H2O ⇌ Ferulate + Polysaccharide. In addition, although for LmjF.36.4640 the function is described in databases as TritrypDB, UniProtKB and GeneDB in the KEGG database, this protein does not have a number of EC assigned.

The EC:3.1.30.1 is responsible for endonucleolytic cleavage to 5'-phosphomononucleotide and 5'-phosphooligonucleotide end-products. However, for the enzymes, LmjF.31.2300 and LmjF.31.2310, termed respectively, putative 3'-nucleotidase/nuclease and putative 3'-nucleotidase/nuclease precursor, there has been no allocation of enzymatic function, or a number of EC, by KEGG database, but this assignment has been made and is available in other databases such as TritrypDB, UniProtKB and GeneDB.

The EC:3.2.1.26 determines the function of five specific enzymes of L. major, LmjF.04.0310, LmjF.04.0320, LmjF.23.0870, LmjF.23.0880 and LmjF.35.0640, these being with activity of beta-fructosidase-like protein involved in Hydrolysis of the terminal non-reducing beta-D-fructofuranoside residues in beta-D- fructofuranosidase in Galactose metabolism and in Starch and sucrose metabolism (lma00052 and lma00500). Although for LmjF.04.0310, LmjF.04.0320 and LmjF.35.0640 the function is described in databases as TritrypDB, UniProtKB and GeneDB in the KEGG database, these enzymes do not yet have a number of EC assigned.

The EC:3.2.2.1 of the specific enzyme putative inosine-adenosine-guanosine-nucleoside
Annexe

hydrolase, LmjF.29.2800, of \textit{L. major} is involved in N-D-Ribosylpurine + H2O $\rightleftharpoons$ Purine + D-Ribose acting in Purine metabolism and Nicotinate and nicotinamide metabolism (lma00230 and lma00760).

The enzymatic function EC:3.4.11.14 annotated for LmjF.17.0140 enzyme has activity on a peptidase putative T, which provides a release of an N-terminal amino acid, preferentially alanine, from a wide range of peptides, amides and arylamines.

The EC:3.4.17.19 designates the release function of the C-terminal amino acid with broad specificity, except for -Pro for two enzymes of \textit{L. major}, LmjF.14.0180 and LmjF.33.2540, termed respectively, the putative carboxypeptidase and metallopeptidase, Clan MA (E) Family M32.


The function of the enzyme acetylornithine deacetylase-like protein of LmjF.07.0270 described by annotation EC: 3.5.1.16 is responsible for the reaction N-Acetylornithine + H2O $\rightleftharpoons$ Acetate + L-Ornithine performed on Arginine biosynthesis (lma00220).

The EC:4.1.1.74 has function of putative pyruvate/indole-pyruvate carboxylase for LmjF.34.3250 enzyme catalyzing the reaction Indolepyruvate $\rightleftharpoons$ indole-3-acetaldehyde + CO2 in Tryptophan metabolism (lma00380).

The EC:4.1.99.3 denotes function for two enzymes of \textit{L. major}, LmjF.09.0360 and LmjF.33.0470, called respectively putative DNA photolyase and putative deoxyribodipyrimidine photolyase. These, in turn, catalyze the reaction Cyclobutadipyrimidine $\rightleftharpoons$ 2 pyrimidine 5'-deoxynucleotide in DNA.

Enzymatic activity EC:4.4.1.8 with function noted for cystathionine beta-lyase-like protein, LmjF.14.0460, catalyzes the reaction L-cystathionine + H2O $\rightleftharpoons$ L-Homocysteine + Ammonia +
Annexe
Pyruvate in Cysteine and methionine metabolism and Selenocompound metabolism (lma00270 and lma00450).

The EC:5.4.99.9 is responsible for inferring function that catalyzes the UDP-alpha-D-galactose <=> UDP-alpha-D-galactofuranose reaction active in galactose metabolism and amino sugar and nucleotide sugar metabolism (ec00052 and ec00520). However, LmjF.18.0200 enzyme, UDP-galactopyranose mutase, has no function assigned in the KEGG database, but has assigned function in other databases such as TritrypDB, UniProtKB and GeneDB.

The enzymatic activity EC:6.2.1.12 has function responsible for the catalysis of ATP + 4-Coumarate + CoA <=> AMP + Diphosphate + p-Coumaroyl-CoA participating in the Ubiquinone and other terpenoid-quinone biosynthesis, Phenylalanine metabolism and phenylpropanoid biosynthesis (ec00130, ec00360 and ec00940). However, LmjF.19.0985, LmjF.19.0995 and LmjF.19.1005 enzymes, 4-coumarate:coa ligase-like protein, have assigned function only in databases as TritrypDB, UniProtKB and GeneDB.

The EC:6.3.1.1 corresponds to the putative function aspartate--ammonia ligase of LmjF.26.0830 enzyme that catalyzes the reaction ATP + L-Aspartate + Ammonia <=> AMP + Diphosphate + L-Asparagine present in Alanine, aspartate and glutamate metabolism and Cyanoamino acid metabolism (lma00250 and lma00460).

The EC:6.3.1.9 is responsible for the function of three enzymes specific putative trypanothione synthetase of L. major, LmjF.23.0460, LmjF.27.1870 and LmjF.36.4300. In addition, two reactions are associated with the EC in the Glutathione metabolism (lma00480). First, to generate Glutathionylspermidine by reaction ATP + Glutathione + Spermidine <=> ADP + Orthophosphate + Glutathionylspermidine. And the second, to generate Trypanothione by reaction ATP + Glutathionylspermidine + Glutathione <=> ADP + Orthophosphate + Trypanothione.

And finally, the enzymatic activity EC:6.5.1.3 with function for two enzymes, LmjF.01.0590 and LmjF.20.1730, called respectively, mitochondrial RNA ligase 2 and mitochondrial RNA editing ligase 1 catalyzes the ATP + RNA(n) + RNA(m) <=> AMP + Diphosphate + RNA(n + m) reaction.
Annexe

S5: Proteins of *T. cruzi*, *T. brucei* and *P. falciparum* homologous to the true positives of *L. major* (*L. major* proteins without any homologies with the human genome).

<table>
<thead>
<tr>
<th>First Comparison</th>
<th>Second Comparison</th>
<th>Third Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>T. cruzi</em></td>
<td><em>L. major</em></td>
<td><em>T. brucei</em></td>
</tr>
<tr>
<td>TcCLB.503745.30</td>
<td>LmjF.34.0070</td>
<td>Tb427.04.0247</td>
</tr>
<tr>
<td>TcCLB.503781.60</td>
<td>LmjF.36.1660</td>
<td>Tb427.01.3030</td>
</tr>
<tr>
<td>TcCLB.504153.160</td>
<td>LmjF.33.2540</td>
<td>Tb427.02.4370</td>
</tr>
<tr>
<td>TcCLB.504191.10</td>
<td>LmjF.36.2380, LmjF.36.2390</td>
<td>Tb427.05.3830</td>
</tr>
<tr>
<td>TcCLB.504191.19</td>
<td>LmjF.36.2380, LmjF.36.2390</td>
<td>Tb427.07.6270</td>
</tr>
<tr>
<td>TcCLB.505683.10</td>
<td>LmjF.36.2380, LmjF.36.2390</td>
<td>Tb427.10.6090</td>
</tr>
<tr>
<td>TcCLB.506193.60</td>
<td>LmjF.34.0070</td>
<td>Tb427.10.6910</td>
</tr>
<tr>
<td>TcCLB.506297.190</td>
<td>LmjF.11.1000</td>
<td>Tb427.10.6950</td>
</tr>
<tr>
<td>TcCLB.506363.110</td>
<td>LmjF.20.1730</td>
<td>Tb427tmp.02.4150</td>
</tr>
<tr>
<td>TcCLB.506401.220</td>
<td>LmjF.18.0200</td>
<td>Tb427tmp.160.2970</td>
</tr>
<tr>
<td>TcCLB.506481.10</td>
<td>LmjF.19.1020</td>
<td>Tb927.1.3030</td>
</tr>
<tr>
<td>TcCLB.506513.110</td>
<td>LmjF.17.0140</td>
<td>Tb927.10.15650</td>
</tr>
<tr>
<td>TcCLB.506975.9</td>
<td>LmjF.20.1730</td>
<td>Tb927.10.6090</td>
</tr>
<tr>
<td>TcCLB.507091.40</td>
<td>LmjF.16.0530</td>
<td>Tb927.10.6910</td>
</tr>
<tr>
<td>TcCLB.507165.50</td>
<td>LmjF.36.3590</td>
<td>Tb927.10.6950</td>
</tr>
<tr>
<td>TcCLB.507639.50</td>
<td>LmjF.36.1660</td>
<td>Tb927.11.6280</td>
</tr>
<tr>
<td>TcCLB.507793.20</td>
<td>LmjF.36.3590</td>
<td>Tb927.2.4370</td>
</tr>
<tr>
<td>TcCLB.508153.640</td>
<td>LmjF.29.2800</td>
<td>Tb927.5.3830</td>
</tr>
<tr>
<td>TcCLB.508375.50</td>
<td>LmjF.16.0530</td>
<td>Tb927.7.6270</td>
</tr>
<tr>
<td>TcCLB.508543.140</td>
<td>LmjF.18.0200</td>
<td>Tb927.9.4360</td>
</tr>
<tr>
<td>TcCLB.508917.49</td>
<td>LmjF.17.0140</td>
<td>Tb927.2.2440</td>
</tr>
<tr>
<td>TcCLB.509027.10</td>
<td>LmjF.09.0360</td>
<td>Tb927.1.1840</td>
</tr>
<tr>
<td>TcCLB.510101.140</td>
<td>LmjF.11.1000</td>
<td>Tb927.10.19090</td>
</tr>
<tr>
<td>TcCLB.510155.20</td>
<td>LmjF.01.0590</td>
<td>Tb927.10.7460</td>
</tr>
<tr>
<td>TcCLB.510185.10</td>
<td>LmjF.36.2380, LmjF.36.2390</td>
<td>Tb927.10.8450</td>
</tr>
<tr>
<td>TcCLB.510315.10</td>
<td>LmjF.29.2800</td>
<td>Tb927.11.7080</td>
</tr>
<tr>
<td>TcCLB.510341.4</td>
<td>LmjF.26.2610</td>
<td>Tb927.5.5320</td>
</tr>
<tr>
<td>TcCLB.511277.240</td>
<td>LmjF.18.0560</td>
<td>Tb927.7.7240</td>
</tr>
<tr>
<td>TcCLB.511295.40</td>
<td>LmjF.19.1020</td>
<td>Tb927.9.2300</td>
</tr>
<tr>
<td>TcCLB.511585.20</td>
<td>LmjF.01.0590</td>
<td>Tb927.10.7460</td>
</tr>
<tr>
<td>TcCLB.511643.20</td>
<td>LmjF.16.0530</td>
<td>Tb927.11.6280</td>
</tr>
<tr>
<td>TcCLB.511923.120</td>
<td>LmjF.16.0530</td>
<td>Tb927.2.4370</td>
</tr>
<tr>
<td>TCSYLVIO_001184</td>
<td>LmjF.16.0530</td>
<td>Tb927.5.3830</td>
</tr>
<tr>
<td>TCSYLVIO_001633</td>
<td>LmjF.11.1000</td>
<td>Tb927.7.6270</td>
</tr>
<tr>
<td>TCSYLVIO_001634</td>
<td>LmjF.11.1000</td>
<td>Tb927.9.4360</td>
</tr>
<tr>
<td>TCSYLVIO_003430</td>
<td>LmjF.01.0590</td>
<td>Tb927.2.2440</td>
</tr>
<tr>
<td>TCSYLVIO_004399</td>
<td>LmjF.34.0070</td>
<td>Tb927.1.1840</td>
</tr>
<tr>
<td>TCSYLVIO_004443</td>
<td>LmjF.29.2800</td>
<td>Tb927.10.19090</td>
</tr>
<tr>
<td>TCSYLVIO_004940</td>
<td>LmjF.36.1660</td>
<td>Tb927.10.7460</td>
</tr>
<tr>
<td>TCSYLVIO_006003</td>
<td>LmjF.19.1020</td>
<td>Tb927.9.2300</td>
</tr>
<tr>
<td>TCSYLVIO_006182</td>
<td>LmjF.18.0560</td>
<td>Tb927.11.6280</td>
</tr>
<tr>
<td>TCSYLVIO_007533</td>
<td>LmjF.36.2380</td>
<td>Tb927.10.6910</td>
</tr>
</tbody>
</table>
Annexe

<table>
<thead>
<tr>
<th>TCSYLVIO_008128</th>
<th>LmjF.36.2390</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCSYLVIO_008411</td>
<td>LmjF.36.3590</td>
</tr>
<tr>
<td>TCSYLVIO_008551</td>
<td>LmjF.20.1730</td>
</tr>
<tr>
<td>TCSYLVIO_009115</td>
<td>LmjF.11.1000</td>
</tr>
<tr>
<td>TCSYLVIO_009755</td>
<td>LmjF.17.0140</td>
</tr>
<tr>
<td>TCSYLVIO_009755</td>
<td>LmjF.18.0200</td>
</tr>
</tbody>
</table>