VINÍCIUS SOARES LAGHI

GP63 DE *TRYPANOSOMA RANGELI*: UM COMPARATIVO *IN-SILICO* DA VARIABILIDADE E ESTRUTURA PROTEÍCA DESTAS METALOPROTEASES COM OUTROS TRIPANOSOMATÍDEOS.

Trabalho de Conclusão de Curso submetido à Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do Grau de Bacharel em Ciências Biológicas. Sob orientação do Professor Glauber Wagner e coorientação da Professora Patrícia Hermes Stoco.

Florianópolis
2016

AGRADECIMENTOS

Aos meus familiares e amigos por terem me apoiado e terem tido paciência comigo durante essa fase difícil da minha vida.

Ao Glauber por ter aceitado ser meu orientador logo que chegou na UFSC, por toda paciência para me ajudar ao longo de todo o período do desenvolvimento desse trabalho e sem o qual não teria conseguido terminar esse projeto.

A Patrícia por ter me introduzido ao laboratório de Bioinformática, por ter me ajudado com o desenvolvimento do projeto do trabalho e por ter me acompanhado no desenvolvimento de todo o trabalho.

Aos membros da banca por terem aceitado meu convite e participados da apresentação desse trabalho que representou uma etapa muito importante da minha graduação.

Ao Nestor por ter me ajudado com toda a parte de modelagem e por ter tido paciência para me ajudar com todos os erros quer surgiram.

Ao Maurício Carvalho por ter me apoiado e me inspirado a terminar esse projeto, sem você todos os problemas e obstáculos teriam ficado muito mais difíceis de superar. Obrigado por fazer parte da minha vida.

RESUMO

Parasitas tripanosomatídeos são responsáveis por causar diversas doenças no mundo todo. O *Trypanosoma rangeli* pode infectar seres humanos e, apesar de não ser patogênico, devido à similaridade antigênica com o *Trypanosoma cruzi*, o seu diagnóstico pode resultar em falsos positivos no diagnóstico sorológico da doença de chagas. A proteína GP63 é uma metaloprotease zinco dependente presente em tripanosomatídeos e associada com a capacidade de infecção desses organismos. Nesse grupo de organismos há uma inserção entre a glicina (G) e a última histidina (H) do motivo HExxHxxGxxH responsável pela ligação ao zinco. Em *T. rangeli* foram detectados três grupos dessa proteína, um contendo o motivo HExxH conservado, outra com uma substituição tornando o motivo em HAxxH e um terceiro grupo totalmente modificado RGxxY. Neste trabalho foi realizado um estudo *in-silico* de sequências de GP63 de *T. rangeli* em comparação com outras sequências de tripanosomatídeos obtidas em banco de dados públicos, além de duas modelagens por homologia de proteínas de *T. rangeli*, uma com o motivo conservado HExxH e outra com o motivo alterado HAxxH afim de analisar possíveis alterações no domínio catalítico. Foi obtido um total de 430 sequências completas de GP63 pertencentes a 27 espécies, classificadas em 17 grupos de ortologia, dos quais apenas quatro grupos possuíam sequências de *T. rangeli*. O modelo criado com o motivo HAxxH revelou uma alteração no domínio devido ao deslocamento da última histidina (H) do motivo, indicando um prejuízo no funcionamento da proteína. A maior parte das sequências apresentou uma inserção perto de 62 aminoácidos, sugerindo que inserções maiores ou menores podem alterar o posicionamento da última histidina (H) como visto nos modelos de *T. rangeli.* A conservação da glicina (G) do motivo e a fenilalanina (F) que a sucede sugere uma importância desses aminoácidos na organização estrutural da proteína. O ácido glutâmico (E) apresentou uma menor frequência em relação aos demais aminoácidos do motivo, sugerindo uma menor importância para o funcionamento da proteína. Diante dos resultados aqui apresentados, recomenda-se a realização de experimentos de bancada que comprovem possíveis alterações na atividade catalítica destas proteínas que possuam uma substituição do ácido glutâmico (E) do motivo ou um deslocamento da última histidina (H) devido ao tamanho da inserção.

Palavras-chave: bioinformática; GP63; *Trypanosoma rangeli*; modelagem por homologia

ABSTRACT

Trypanosomatids are responsible for causing many diseases around the world. *Trypanosoma rangeli* can infect humans and, although it's not pathogenic, due to its antigenic similarity with *Trypanosoma cruzi*, its diagnosis can result in false positives in the serological diagnosis of the chagas disease. The GP63 protein is a zinc dependent metalloprotease present in trypanosomatids and related to their infection capacity. In these organisms there is an insertion between the glycine (G) and the last histidine (H) of the motif HExxHxxGxxH responsible for bonding with the zinc. In *T. rangeli* three groups of this protein were detected, one containing the conserved HExxH motif, another one with a substitution that changes the motif to HAxxH and a third one completely modified RGxxY. In this work, a in silico study was carried out regarding *T. rangeli* sequences compared to other trypanosomatids sequences obtained from public databases, in addition to two homology modeling of *T. rangeli* proteins, one containing the conserved motif HExxH and another one with the changed motif HAxxH, in order to analyze possible alterations in the catalytic domain. A total of 430 complete sequences were obtained, belonging to 27 species, classified in 17 orthology groups, of which only four contained *T. rangeli* sequences. The model made with the HAxxH motif revealed a change in the catalytic domain, due to the displacement of the last histidine (H) of the motif, indicating a loss in the protein activity. Most sequences presented an insertion around 62 amino acids, suggesting that bigger or smaller insertions could result in a change of the positioning of the last histidine (H), as seen in the *T. rangeli* protein models. The conservation of the glycine (G) and the phenylalanine (F) in the motif suggests an importance of these amino acids for the structural organization of the protein. The glutamic acid (E) was the least frequent in relation to the other amino acids of the motif, suggesting a smaller importance for the protein functioning. Because of the results presented here, it would be recommended to perform further experiments to test for any possible alterations in the catalytic activity of these proteins that show substitution of the glutamic acid (E) or displacement of the last histidine (H) due to the size of the insertion.

Keywords: bioinformatics; GP63; *Trypanosoma rangeli*; homology modeling.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BLAST	*Basic Align Search Tool* – Ferramenta de Pesquisa de Alinhamento Básico
CDD	*Conserved Domain Database* – Banco de Dados de Domínios Conservados
DNA	*Deoxyribonucleic Acid* – Ácido desoxirribonucléico
GPI	Glicosilfosfatidilinositol
kDa	Kilodalton
NCBI	*National Center for Biotechnology Information* – Centro Nacional de Informação Biotecnológica.
PDB	*Protein Data Bank* – Banco de Dados de Proteína

SUMÁRIO

1 INTRODUÇÃO

1.1 TRIPANOSOMATÍDEOS

Parasitos tripanosomatídeos são protozoários pertencentes à família Trypanosomatidae, caracterizados por serem parasitos obrigatórios, unicelulares e flagelados (ASLETT, AURRECOECHEA, *et al.*, 2010) e pertencentes à ordem Kinetoplastida que tem como característica a presença de uma região rica em DNA extracromossomal chamada cinetoplasto (kDNA), localizadas na base do flagelo. Dentro desse grupo há dois gêneros causadores de doenças em seres humanos, o gênero *Trypanosoma* e o *Leishmania* (NEVES, MELO, *et al.*, 2011), que afetam diversas pessoas em todo o mundo (BANGS, RANSOM, *et al.*, 2001). Dentre as doenças causadas por esses gêneros podemos citar a Doença de Chagas nas Américas (*Trypanosoma cruzi*), a Doença do Sono na África (*Trypanosoma brucei*) e a Leishmaniose cutânea e visceral em todo o mundo (*Leishmania* spp.) (ASLETT, AURRECOECHEA, *et al.*, 2010).Os demais gêneros de tripanosomatídeos parasitam invertebrados, principalmente insetos, com exceção do gênero Phytomonas que parasita plantas e do gênero *Endotrypanum* que também parasita vertebrados (NEVES, MELO, *et al.*, 2011).

Os tripanosomatídeos podem possuir um ciclo monoxeno ou heteroxeno variando de espécie pra espécie, apresentando alternâncias de formas celulares durante o ciclo (NEVES, MELO, *et al.*, 2011). As diferentes formas celulares possuem variações quanto à posição, inserção e tamanho do flagelo, a presença e a forma de uma membrana ondulante, a posição do cinetoplasto e a forma celular, e são nomeadas de acordo com essas características (NEVES, MELO, *et al.*, 2011).

1.2 Trypanosoma rangeli

O *Trypanosoma rangeli* (Tejera, 1920) possui um ciclo de vida heteroxeno, tendo como vetor insetos triatomíneos dos gêneros *Rhodnius* e *Triatoma*, e podendo infectar diversos mamíferos de até cinco diferentes ordens (NEVES, MELO, *et al.*, 2011; STOCO, WAGNER, *et al.*, 2014). O ciclo inicia quando um triatomíneo ingere as formas tripomastigotas sanguíneas deste parasita durante o repasto sanguíneo em um mamífero infectado. Após a ingestão, o parasita se diferencia em epimastigotas se multiplica e migra para as glândulas salivares do inseto, onde se diferencia novamente em formas

tripomastigotas metacíclicas. Por fim estas formas são inoculadas no hospedeiro mamífero durante um novo repasto sanguíneo (NEVES, MELO, *et al.*, 2011; STOCO, WAGNER, *et al.*, 2014).

O *T. rangeli*, assim como o *T. cruzi* e o *T. brucei*, tem a capacidade de infectar seres humanos, entretanto, ao contrário das demais espécies, o *T. rangeli* não é patogênico para seus hospedeiros mamíferos, e sim para seu vetor invertebrado. Apesar de não patogênico, em função da similaridade antigênica com o *T. cruzi* o diagnóstico do *T. rangeli* pode acarretar em resultados falsos positivos no diagnóstico sorológico da doença de chagas em função da reatividade cruzada. Além disso, o *T. rangeli* ocorre em quase todos os países da América do Sul e Central, sobrepondo a região de ocorrência de *T. cruzi*, dificultando ainda mais o diagnóstico adequado (NEVES, MELO, *et al.*, 2011; STOCO, WAGNER, *et al.*, 2014).

1.3 GP63

O ciclo de vida de tripanosomatídeos envolve uma série de adaptações que os possibilitam parasitarem seus hospedeiros, como uma alternância de formas celulares, proteínas e estruturas relacionadas com a capacidade de infecção e patogênese do parasito (NEVES, MELO, *et al.*, 2011). Uma dessas proteínas envolvidas com a infectividade e a patogênese desses parasitos é a GP63, também chamada de *Leishmanolysin* ou MSP (*Major Surface Protease*).

A proteína GP63 é uma metaloprotease zinco dependente, com peso de 60-65 kDa e com uma âncora de GPI (PEREIRA, 2014). Em *Leishmania* sp. e *T. cruzi*, é responsável por auxiliar na infecção celular e na proteção contra degradação lisossomal desses parasitas (CHANG e CHANG, 1986; MCGWIRE e CHANG, 1994; KULKARNI, OLSON *et al.*, 2009). Nos estágios de vida de *T. brucei* presentes na corrente sanguínea do hospedeiro, essa proteína é responsável por desprender as glicoproteínas variáveis de superfície (VSGs) desse parasita (LACOUNT, GRUSZYNSKI, *et al.*, 2003). Apesar de serem relacionadas com a capacidade infecciosa desses e outros parasitos dessa família em mamíferos, a GP63 também está presente em espécies não-patogênicas como *Phytomonas* sp., *Herpetomonas* sp. (SANTOS, BRANQUINHA e D'VILLA-LEVY, 2006) e *T. rangeli* (FERREIRA, RUIZ, *et al.*, 2010;GRISARD, STOCO, *et al.*, 2010). Estudos sugerem que no inseto vetor essa proteína está associada com a adesão do parasito no intestino do inseto (SANTOS, BRANQUINHA e D'VILLA-LEVY, 2006; YAO, 2010) além de supostamente estar associada com a

nutrição do parasito dentro do vetor (SANTOS, BRANQUINHA e D'VILLA-LEVY, 2006)

Quando expressada *in-vitro*, a GP63 é capaz de hidrolisar diversos substratos protéicos como albumina, componentes do sistema complemento, hemoglobina, fibrinogênio e imunoglobulinas, além de componentes da matriz extracelular como colágeno tipo IV e fibronectina (PEREIRA, 2014). Entretanto, ainda não se tem total conhecimento dos substratos dessa proteína *in-vivo* (YAO, DONELSON e WILSON, 2003; PEREIRA, 2014)

Em *T. cruzi* foram identificados dois grupos de genes para a família GP63 (TCGP63-1 e TCGP63-2), ambos com os resíduos de histidina e ácido glutâmico conservados no motivo HExxH, que são os resíduos mais importantes para a atividade catalítica (CUEVAS, CAZZULO e SÁNCHEZ, 2003). Esse motivo está diretamente associado com o domínio catalítico da proteína, sendo responsável por realizar a ligação ao zinco catalítico (SCHALANGENHAUF, ETGES e METCALF, 1998)

Em *T. rangeli*, foram identificados três grupos, entretanto somente o grupo GP63-1 de *T. rangeli* apresenta conservação dos resíduos de Histidina e Ácido glutâmico no motivo HExxH, enquanto o grupo GP63-2 apresenta uma substituição do Ácido glutâmico por uma Alanina (HAxxH), o que sugere uma alteração na atividade catalítica e o grupo GP63-3 apresenta um domínio totalmente modificado (RGxxY) (PEREIRA, 2014).

SCHALANGENHAUF, ETGES e METCALF (1998) classificaram uma GP63 de *Leishmania major* como pertencente a família *metzincin*, por possuir um motivo extendido típico dessa família (HExxHxxGxxH), entretanto com a diferença de possuir uma inserção de 62 aminoácidos entre a glicina e a última histidina do motivo, transformando o motivo em $HExxHxxGx_{62}H$.

Embora bem descrita em algumas espécies, ainda existem muitas dúvidas quanto à função dessa proteína em várias espécies de tripanosomatídeos.

1.4 BIOINFORMÁTICA

A bioinformática é uma ciência interdisciplinar na qual faz a interface das ciências da vida e da informação. De maneira simplifica, ela pode ser subdividida em três grandes áreas: análise de sequências de nucleotídeos e aminoácidos, análise de modelos tridimensionais de

proteínas e desenvolvimento de aplicações para análise de dados biológicos (VERLI, 2014).

Com o desenvolvimento de técnicas de sequenciamento de DNA/RNA desde o projeto Genoma Humano, houve uma produção crescente de dados biológicos (Figura 1) que são gerados em uma velocidade muito superior a velocidade com que eles podem ser trabalhados (VERLI, 2014). Além disto, a maior parte desses dados são depositados em bancos de dados para uso público, como o do National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/) e o TriTrypDB (ASLETT, AURRECOECHEA, *et al.*, 2010).

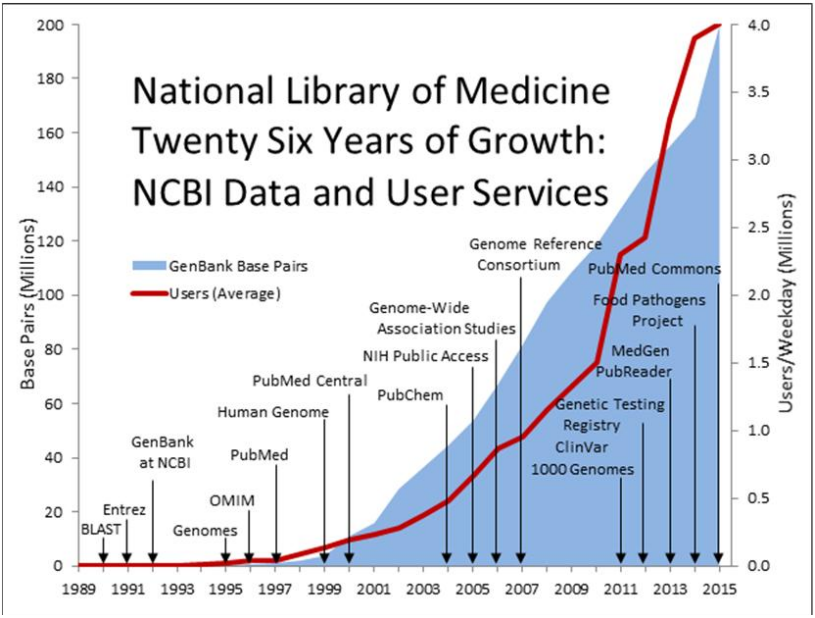

Figura 1: Crescimento do número de usuários e da quantidade de dados depositádos no banco de sequências do NCBI. Note o aumento no crescimento a partir do início do século XXI (Fonte: https://www.nlm.nih.gov/about/2017CJ.html)

1.4.1 ANOTAÇÕES AUTOMÁTICAS DO GENOMA

Após o seqüenciamento do genoma ocorre a identificação de regiões codificantes, que são os genes propriamente ditos, feita através de algoritmos que buscam sinais transcricionais nas sequências que indicam a presença de um gene na região, que em eucariotos se inicia pela região TATA box e termina no sítio de clivagem e poliadenilação (VERLI, 2014). O passo seguinte a identificação dos genes é a anotação da função do gene, feita através de algoritmos de comparações com sequências gênicas ou domínios protéicos conservados depositados em bancos de dados públicos. Primeiramente é feita uma busca por genes ortólogos diretos, e caso encontrado algum, a anotação da sequência será a mesma do gene encontrado. Caso não encontrado nenhum ortólogo, procura-se em seguida por domínios proteícos conservados e a anotação é feita descrevendo os domínios encontrados como "proteína contendo o domínio". Por fim, caso não seja encontrado nada a proteína é descrita como proteína hipotética (VERLI, 2014). É importante lembrar que genes ortólogos são genes que originaram-se em uma espécie ancestral e que se mantiveram nas espécies descendentes, enquanto genes parálogos são aqueles originados de uma duplicação gênica em um mesmo genoma e muitas vezes acabam por divergir em suas funções (ALBERTS, JOHNSON, *et al.*, 2002).

1.4.2 MODELAGEM DE ESTRUTURAS PROTEÍCAS

As proteínas são formadas através de estruturas primárias (sequências de aminoácidos) que se enovelam para adotar uma estrutura secundária que por sua vez formam as estruturas terciárias. Diferentes sequências de aminoácidos (estruturas primárias) podem levar a mesma estrutura terciária, desde que haja a conservação de determinados aminoácidos chaves que desempenham um papel importante na formação da estrutura. Dessa forma, proteínas que tenham um nível de conservação baixo entre suas sequências aminoacídicas podem acabar por apresentar estruturas tridimensionais similares (VERLI, 2014).

Os motivos são pequenas estruturas características de um grupo ou família de proteínas, conservado nas mesmas, que pode ser uma sequência aminoacídica ou uma estrutura secundária, podendo ou não estar associada a alguma atividade catalíticas (BERGERON, 2002; LESK, 2002), enquanto os domínios são regiões estruturais que desempenham um papel quase que independente do resto da proteína (LESK, 2002), que podem ser por exemplo, um sítio de ligação ou um

sítio catalítico. É comum encontrar motivos dentro de um domínio, mas o contrário não acontece.

A modelagem por homologia é uma técnica utilizada para prever a estrutura tridimensional de uma proteína a partir da sua sequência aminoacídica. Esse método consiste em utilizar-se de uma estrutura de uma proteína molde que tenha sido previamente resolvida experimentalmente, que seja similar a proteína alvo e que preferencialmente desempenhe a mesma função ou uma função similar. Quanto maior o grau de similaridade entre a proteína alvo e a proteína molde, melhor o resultado da modelagem. Quando disponível, pode-se utilizar de múltiplos moldes para a modelagem (VERLI, 2014).

Este trabalho utilizou-se de sequências depositadas em bancos de dados públicos para analisar essas proteínas, através de ferramentas de análises de sequências e de predição de estruturas por homologia, a fim de melhor compreender o seu funcionamento em *T. rangeli*.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Determinar a relação entre os diferentes motivos de proteínas GP63 de tripanosomatídeos com sua função, a partir de análises e comparações *in silico* entre sequências de diferentes espécies dessa família, com enfoque em *Trypanosoma rangeli*.

2.2 OBJETIVOS ESPECÍFICOS

- Realizar uma análise do motivo de GP63 de sequências de tripanosomatídeos;
- Classificar as sequências obtidas em grupos de ortologia;
- Realizar uma modelagem por homologia com sequências de diferentes motivos de *T. rangeli*.

3 METODOLOGIA

3.1 RESUMO DA METODOLOGIA

As análises desse trabalho foram todas realizadas *in-silico*, através de ferramentas de análise gratuitas e *scripts* escritos em linguagem Perl. O fluxograma abaixo (Figura 2) mostra uma visão geral de todas as etapas realizadas nesse estudo.

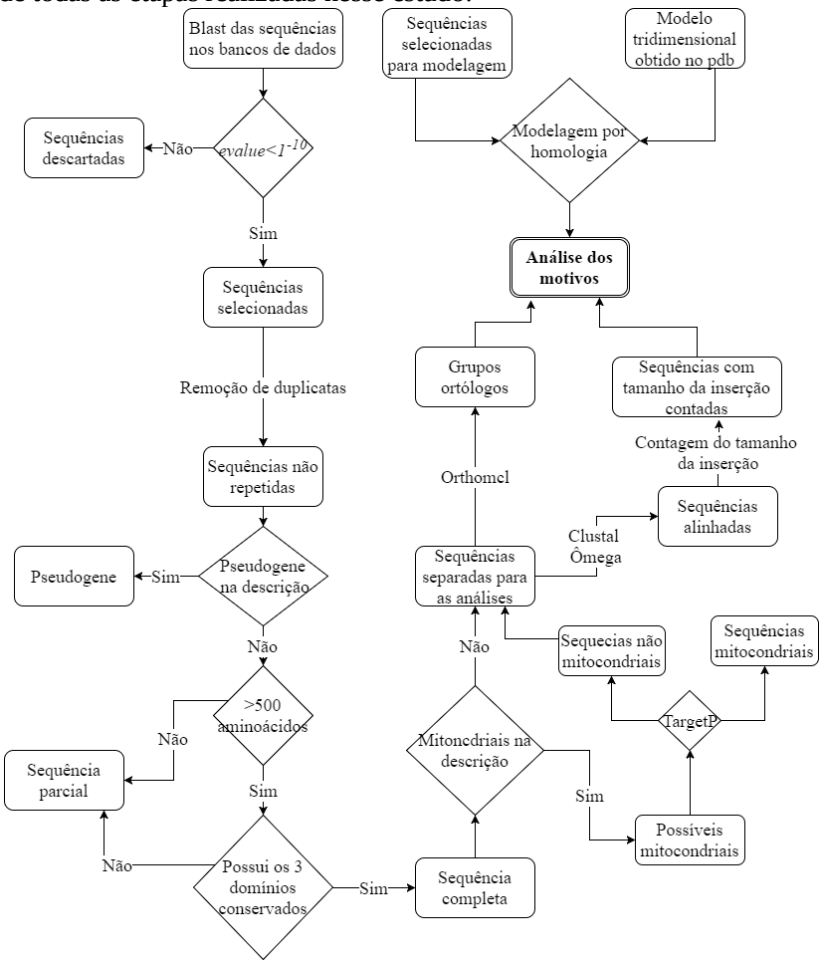

Figura 2: Fluxograma com as etapas realizadas na metodologia (Feito com a ferramenta Draw.io, disponível em: https://www.draw.io/).

3.2 OBTENÇÃO DAS SEQUÊNCIAS

Para a obtenção das sequências utilizadas na análise, foram utilizadas três sequências de GP63 de *T. rangeli* previamente obtidas no trabalho de PEREIRA (2014) como modelos, contendo o motivo HExxH conservado, outra com uma substituição do ácido glutâmico por uma alanina (HAxxH), e uma terceira com o motivo totalmente alterado (RGxxY), determinadas como GP63_A, GP63_B e GP63_C, respectivamente. Com estas sequencias foram realizados análises de similaridade utilizando o programa BLAST com os bancos de dados de proteínas dos genomas depositados no TriTrypDB (http://tritrypdb.org/tritrypdb/) e o banco de dados não redundante (NR) do NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi), utilizando como parâmetros de referência valores de $e < 1^{-10}$.

3.3 TRATAMENTO DAS SEQUÊNCIAS OBTIDAS

Em função da existência de sequências duplicadas ou com alta similaridade (maior que 99%) entre as sequências obtidas em ambos os bancos de dados, algumas sequências eram repetidas. Para a remoção desses duplicados foi escrito um *script* na linguagem Perl, executado nos dados obtidos do NCBI e do TriTrypDB separadamente, gerando dois arquivos com as sequências não redundantes para cada banco. Depois, as sequências interbancos foram comparadas, objetivando identificar as sequências repetidas e aquelas que deram um match de 100% de identidade e possuíam o mesmo tamanho foram consideradas repetidas, permanecendo apenas uma sequencia no conjunto final de dados.

Para identificar a presença de domínios característicos da família das metaloproteases, foi utilizado o programa RPS-Blast contra o banco de dados CDD v.3.14 (maio de 2015) considerando valor de e de $< e^{-10}$.

As sequências não redundantes foram classificadas em: proteínas completas caso possuíssem uma sequência maior que 500 aminoácidos e a presença dos três domínios característicos dessa classe de proteínas (PFAM01457, PTZ00337 e PTZ00257) parciais caso fossem menor que 500 aminoácidos com ao menos um domínio ou maiores que 500 aminoácidos, mas sem a presença dos três domínios; e por último pseudogenes, caso assim estivessem classificadas em sua descrição original. As sequências que não possuíam nenhum dos três domínios característicos foram descartadas. A determinação do tamanho de 500 aminoácidos utilizada para a classificação entre sequências completas e

parciais foi um critério arbitrário em virtude da grande quantidade de sequências curtas que foram obtidas.

3.4 IDENTIFICAÇÃO DE ORTÓLOGOS

O programa utilizado para a identificação de grupos ortólogos foi o Orthomcl v.2.0 (LI, CHRISTIAN, *et al.*, 2003) Para tal, as sequencias não redundantes foram classificadas de acordo com a espécie e cepa, quando existente.

3.5 ANÁLISE DO MOTIVO

Para a análise do motivo das sequências completas foi feito um alinhamento com a ferramenta Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/), para então editar as sequências com o programa BioEdit (HALL, 1999). No programa BioEdit foram identificados o início e o fim do motivo completo ($HxxxHxxGx_{(n)}H$) utilizando-se como base a sequência do cristal descrito por SCHLAGENHAUF *et al*., 1998 (PDB: 1LML). Após a identificação dos grupos das sequências, as sequências foram tratadas mantendo apenas a região do motivo e exportadas em formato fasta para serem analisadas no programa WebLogo (CROOKS, HON, *et al.*, 2004), para gerar as imagens com os motivos a serem analisadas. Além disso, foi feito um *script* para identificar o tamanho da inserção entre a última Glicina (G) do motivo e uma histidina (H) que foi observada como comum na maioria das sequências analisadas.

3.6 MODELAGEM

Foram selecionadas duas sequências para a realização de modelos tridimensionais para a análise dos motivos. A primeira sequência selecionada para a modelagem foi a sequência GP63_B, utilizada como modelo para a realização dos BLASTs com o domínio HAxxH. A segunda sequência foi selecionada a partir dos resultados de ortologia, onde se escolheu uma sequência de *T. rangeli* com o motivo HExxH (NCBI: ESL05225.1), pertencente à um grupo de ortologia compartilhado com sequências de *T. cruzi*. Para a realização da modelagem foi utilizado o programa I-TASSER v.4.4 (YANG, YAN, *et al.*, 2015), utilizando como base um cristal de GP63 de *Leishmania major* descrito por SCHLAGENHAUF et al. (1998), depositado no Protein Data Bank (PDB) (BERMAN, WESTBROOK, *et al.*, 2000) sob

o código 1LML, com as configurações no *default*. A proteína utilizada como base no modelo não possuía os peptídeos sinais em seu cristal e, portanto, com o intuito de obter melhores resultados na modelagem, as sequências a serem utilizadas tiveram os mesmos removidos, utilizando a ferramenta Clustal Omega Online (http://www.ebi.ac.uk/Tools/msa/clustalo/) para alinhá-la com a sequência modelo (1LML) para identificar o ponto onde a sequência deveria ser cortada.

3.7 VERIFICAÇÃO DE SEQUÊNCIAS MITOCONDRIAIS

As sequências anotadas como mitocondriais foram identificadas e separadas em um arquivo fasta, para então serem analisadas através da ferramenta TargetP v1.1 (EMANUELSSON, NIELSEN, *et al.*, 2000).

4 RESULTADOS

4.1 SEQUÊNCIAS ANALISADAS

Foi obtido um total de 1497 sequências, das quais 430 foram classificadas como sequências completas, pertencentes a cinco gêneros e 27 espécies de kinteoplastideos (Tabela 1), enquanto as demais foram classificadas como pseudogenes ou parciais. Além das sequências obtidas dos bancos de dados, foram adicionadas manualmente duas sequências utilizadas para a geração do modelo tridimensional: a sequência utilizada para a criação do modelo tridimensional, pertencente a *Trypanosoma rangeli*, que também foi utilizada como um dos modelos para a execução dos BLASTs (GP63_B) e a sequência do cristal utilizado como base para a geração do modelo, pertencente à *Leishmania major*, sob o código 1LML no PDB, totalizando 432 sequências.

4.2 MODELO TRIDIMENSIONAL

O primeiro modelo gerado pelo programa, da sequência GP63_B (modelo 1) de *T. rangeli* apresentou um *score* de 1.18 e uma grande similaridade com o cristal utilizado como base (modelo 1LML). É possível perceber uma sobreposição das estruturas secundárias α hélice quase que perfeitamente (Figura 3), além do posicionamento do domínio catalítico (motivo $HxxxHxxGx_nH$) ser extremamente similar, com exceção da última histidina (H) posterior a inserção X_n (Figura 4). A análise da carga da superfície revelou várias regiões com carga positiva para o modelo 1, enquanto o modelo 1LML apresenta uma superfície predominantemente negativa (Figura 5).

O segundo modelo (modelo 2) gerado com a sequência NCBI: ESL05225.1 também de *T. rangeli* apresentou um *score* de 1.47 e uma similaridade ainda maior com o cristal do modelo 1LML (Figura 6). Além de não haver a substituição do ácido glutâmico (E) por uma alanina (A) no motivo do domínio catalítico, a histidina H334 posterior a inserção também está alinhada com a do cristal (Figura 7).

Tabela 1: Total de sequências obtidas por espécie que foram utilizadas nas análises subsequentes.

Espécie	Sequências obtidas
Crithidia fasciculata	18
Endotrypanum monterogeii	53
Leishmania aethiopica	2
Leishmania amazonensis	1
Leishmania arabica	2
Leishmania braziliensis	30
Leishmania donovani	11
Leishmania enriettii	26
Leishmania gerbilli	2
Leishmania guyanensis	3
Leishmania infantum	8
Leishmania major	14
Leishmania mexicana	8
Leishmania panamensis	4
Leishmania sp.	4
Leishmania tarentolae	5
Leishmania tropica	2
Leishmania turanica	3
Leptomonas pyrrhocoris	18
Leptomonas seymouri	2
Trypanosoma brucei	29
Trypanosoma congolense	6
Trypanosoma cruzi	138
Trypanosoma evansi	5
Trypanosoma grayi	17
Trypanosoma rangeli	9
Trypanosoma vivax	10

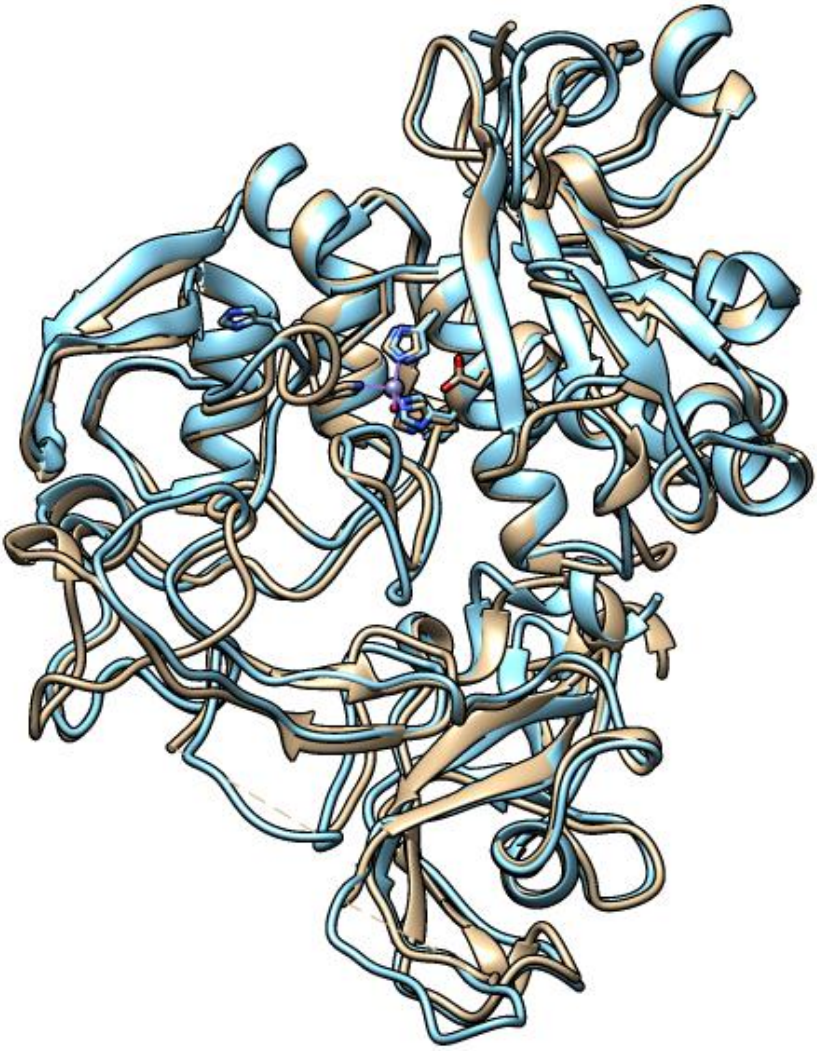

Figura 3: Alinhamento 3D em fita do modelo 1 de GP63 gerado pela modelagem por homologia (em azul) e o modelo 1LML utilizado como base, baseado no cristal de *Leishmania major* (em bege).

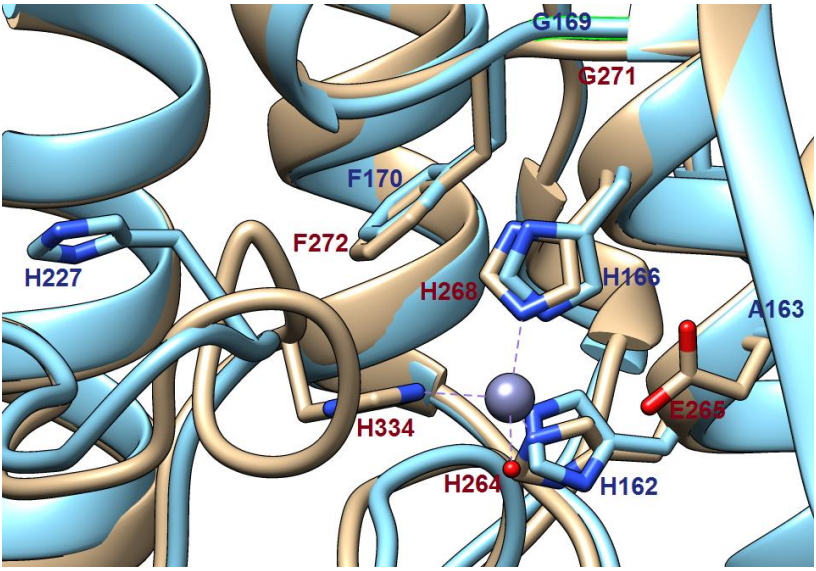

Figura 4: Alinhamento 3D em fita do motivo do modelo 1 (em azul) e o modelo 1LML (em bege). A esfera cinza representa o zinco e a esfera vermelha uma molécula de água.

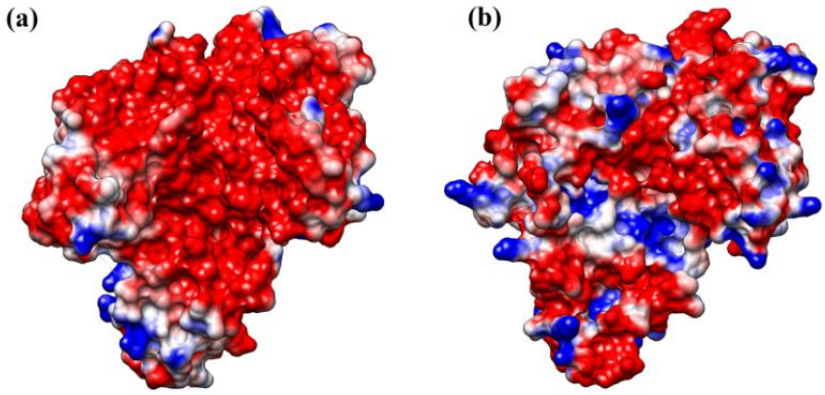

Figura 5: Representação da superfície dos modelos, com coloração em vermelho representando regiões com carga negativa e azul para carga positiva. (a) Modelo 1LML. (b) Modelo 1 (GP63_B).

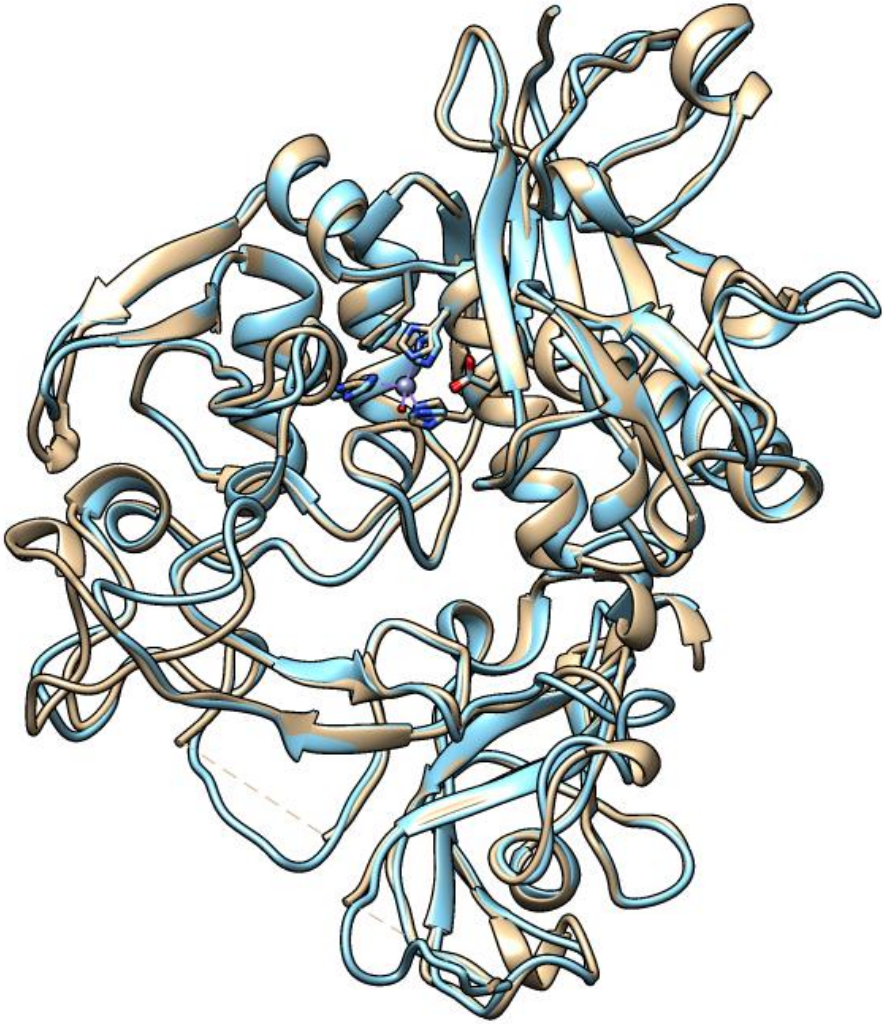

Figura 6: Alinhamento 3D da representação em fita do modelo 2 de GP63 gerado pela modelagem por homologia (em azul) e o modelo 1LML utilizado como base, baseado no cristal de Leishmania major (em bege).

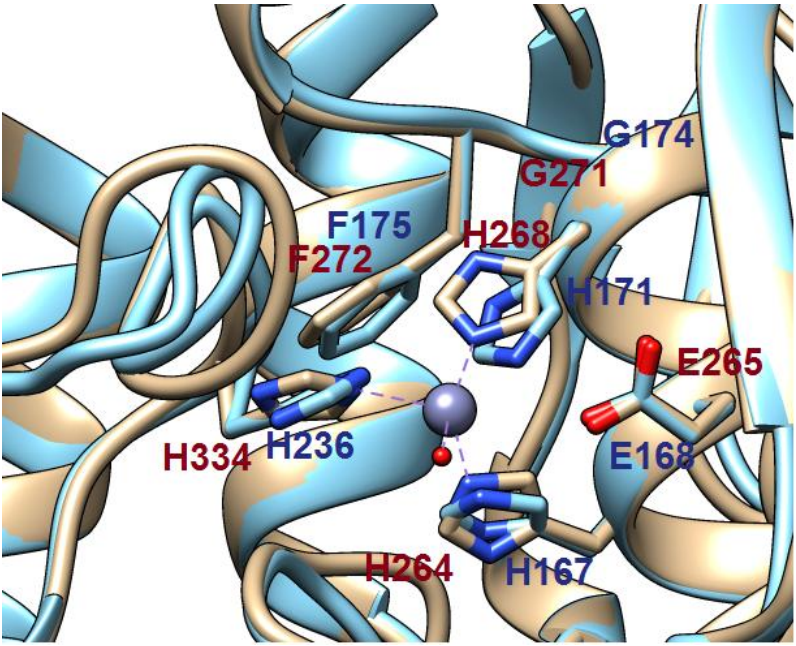

Figura 7: Alinhamento 3D da representação em fita do motivo do modelo 2 (em azul) e o modelo 1LML (em bege). A esfera cinza representa o zinco e a esfera vermelha uma molécula de água.

4.3 ANÁLISE DOS MOTIVOS ALINHADOS

O alinhamento dos motivos de todas as sequências completas revelou uma grande conservação dos aminoácidos pertencentes ao motivo $HxxxHxxGx_nH$ (Figura 8), com uma predominância do ácido glutâmico (E) na posição 2 na maior parte das sequências. Além disso, houve uma conservação muito grande também na fenilalanina (F) que sucede a glicina (G) do motivo na posição 9.

O tamanho da inserção presente entre a glicina e a última histidina variou entre 57 e 289, entretanto a grande maioria das proteínas o tamanho permaneceu perto de 62 resíduos (Figura 9).

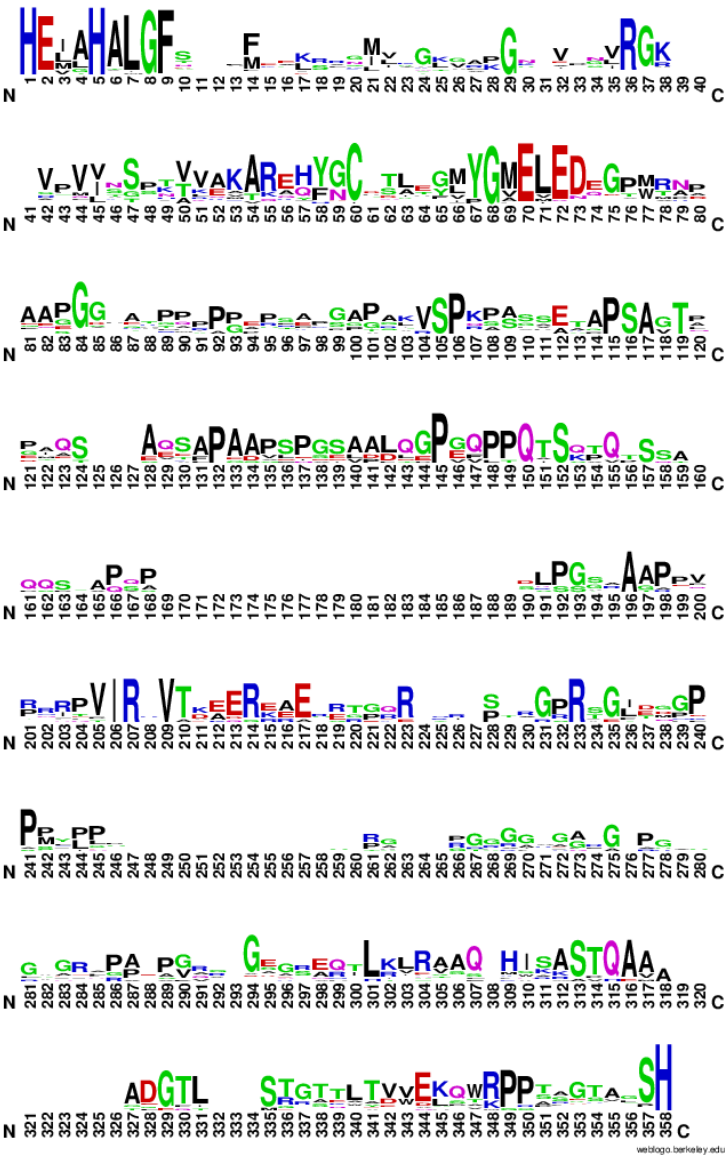

Figura 8: Representação gráfica (Logo) da região do motivo de todas as sequências completas alinhadas no programa Clustal Omega. Quanto maior o tamanho da letra, maior a frequência observada do aminoácido no alinhamento, indicando o grau de conservação do resíduo. Note o grande grau de conservação nos aminoácidos das posições 1, 5, 8, 9 e 358.

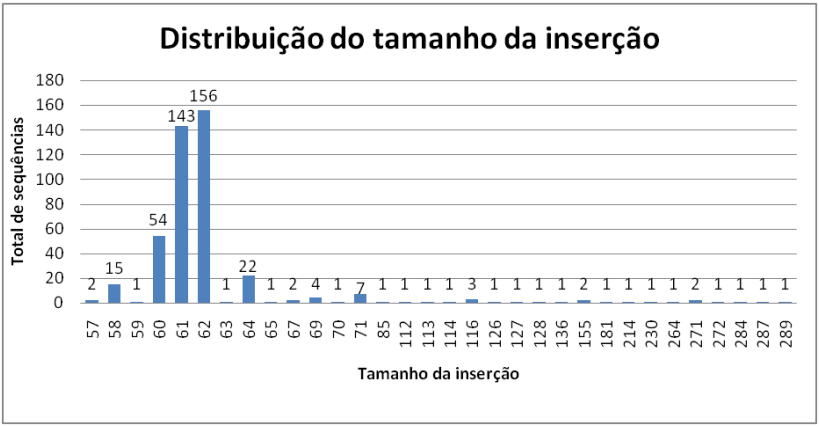

Figura 9: Gráfico da distribuição do tamanho da inserção. Note que a maior parte das sequências ficou entre 61 e 62 aminoácidos.

4.4 GRUPOS DE ORTOLOGIA

Foram obtidos 17 grupos de ortologia, sendo que apenas 4 grupos possuíam sequências de *T. rangeli* (Tabela 2). Para este trabalho descreveremos com detalhes os grupos que possuíam sequências de *T. rangeli*.

O grupo 1 foi o grupo que apresentou a maior quantidade de sequências, com um total de 279 sequências, contemplando 25 espécies (Tabela 3). A variação do tamanho da inserção entre a glicina (G) e a última histidina (H) ficou entre 58 e 62, e somente 12 sequências possuíam uma substituição do ácido glutâmico (E) por uma glutamina (Q) ou aspargina (N) (HxxxH), enquanto as demais apresentaram uma conservação do ácido glutâmico (HExxH).

O grupo 2 é composto por 38 sequências, sendo 36 sequências de *T. cruzi* e 2 sequências de *T. rangeli*. Em todas com o ácido glutâmico conservado no motivo HExxH e a inserção entre a glicina (G) e a ultima histidina (H) do motivo variando entre 61 e 62 aminoácidos. A sequência ESL05225.1 de *T. rangeli* utilizada para a produção do modelo 2 foi classificada nesse grupo.

O grupo 3 é composto por 12 sequências, sendo 11 delas de *T. cruzi* e 1 delas de *T. rangeli*. Todos possuíam o ácido glutâmico (E) do motivo HExxH conservado e a inserção variou entre 155 e 289 aminoácidos entre a glicina e a última histidina do motivo.

O grupo 4 apresentou um total de seis sequências, sendo quatro de *T. cruzi* e duas de *T. rangeli*. Todas essas sequências possuíam uma substituição do ácido glutâmico (E) por outro aminoácido no motivo HxxxH e a inserção entre a glicina e a histidina foi de 57 e 58 para as sequências de *T. rangeli* e 61 aminoácidos para as de *T. cruzi*. A sequência GP63_B utilizada para gerar o modelo tridimensional 1 (Figura 3) foi classificada nesse grupo.

A sequência 1LML não foi utilizada na identificação dos grupos de ortologia por não possuir os peptídeos sinais em sua sequência, entretanto uma análise por BLAST com uma sequência de *T. rangeli* proveniente de cada um dos quatro grupos revelou uma maior similaridade dessa sequência com o grupo de ortologia 1.

Tabela 2: Grupos de ortologia contendo espécies de *T. rangeli*

Grupo de ortologia	Sequências	Espécies	Inserção	Motivo	Modelo
Grupo 1	279	25 spp.	58 a 62	HExxH (12 HxxxH)	1LML
Grupo 2	38	*T. cruzi*; *T. rangeli*	61 a 62	HExxH	Modelo 2
Grupo 3	12	*T. cruzi*; *T. rangeli*	155 a 289	HExxH	N/A
Grupo 4	6	*T. cruzi*; *T. rangeli*	57 a 58	HxxxH	Modelo 1

Tabela 3: Quantidade de sequências pertencentes ao grupo de ortologia 1 por espécie.

Grupo 1	
Espécie	**Sequências**
Crithidia fasciculata	15
Endotrypanum monterogeii	52
Leishmania aethiopica	1
Leishmania amazonensis	1
Leishmania braziliensis	28
Leishmania donovani	10
Leishmania enriettii	25
Leishmania gerbilli	1
Leishmania guyanensis	3
Leishmania infantum	7
Leishmania major	11
Leishmania mexicana	1
Leishmania panamensis	1
Leishmania sp.	3
Leishmania tarentolae	2
Leishmania tropica	1
Leishmania turanica	1
Leptomonas pyrrhocoris	8
Trypanosoma brucei	20
Trypanosoma congolense	3
Trypanosoma cruzi	68
Trypanosoma evansi	4
Trypanosoma grayi	3
Trypanosoma rangeli	5
Trypanosoma vivax	5
Total	279

5 DISCUSSÃO

De todos os gêneros encontrados com sequências consideradas completas, apenas os gêneros *Leptomonas* e *Crithidia* são parasitas monóxenos exclusivos de invertebrados, enquanto os demais gêneros são parasitas heteróxenos de mamíferos. Dentre os gêneros que parasitam mamíferos, temos o *Endotrypanum* que parasita exclusivamente bicho-preguiça e *Trypanosoma* e *Leishmania*, que parasitam humanos dentre outros mamíferos (NEVES, MELO, *et al.*, 2011). Estudos sugerem que no gênero *Crithidia* e no gênero *Leptomonas* a GP63 exerce uma função no processo do parasita se prender ao intestino do inseto (YAO, 2010).

A grande quantidade de sequências de *Leishmania spp. T. brucei e T cruzi* provavelmente reflete o fato de que as espécies de interesse médico são muito mais estudadas e, portanto, possuem mais genomas sequênciados e devidamente anotados (BERRIMAN, GHEDIN, *et al.*, 2005; EL-SAYED, MYLER, *et al.*, 2005; IVENS, PEACOCK, *et al.*, 2005). Além disso há o fato de que essas espécies apresentam várias cópias dessas sequências em seus genomas, chegando a ultrapassar 350 cópias (incluindo pseudogenes) de genes "GP63-*like*" em *T. cruzi* (IVENS, PEACOCK, *et al.*, 2005). Em algumas espécies de parasitas do gênero *Phytomonas* foram encontradas atividades de metaloproteases e em algumas foram encontrados epítopos comuns aos de GP63 de *Leishmania* (YAO, 2010). Entretanto essas espécies não foram encontradas neste estudo, possivelmente por falta da anotação das proteínas em seus genomas sequenciados ou talvez essas proteínas não são GP63 clássicas e, portanto, não foram obtidas no resultado da busca com o programa BLAST.

O motivo catalítico das GP63 apresentou uma conservação das três histidinas (H) (posições 1, 5 e 368) e da glicina (G) (posição 8) do motivo em todas as sequências obtidas, além de uma alta frequência do ácido glutâmico (E) (posição 2) na maior parte das sequências. Nas metaloproteases de zinco que apresentam o motivo HExxHxxGxxH as 3 histidinas (H) se ligam ao zinco, sendo que a glicina (G) é responsável por realizar uma curvatura na estrutura da molécula permitindo que a última histidina do motivo esteja alinhada para poder se ligar ao zinco (SCHALANGENHAUF, ETGES e METCALF, 1998). Apesar da inserção presente em tripanosomatídeos, essa curvatura aparenta ser importante para a conformação da proteína, visto a conservação da glicina (G) nas sequências observadas (Figura 8). Além disso, foi possível observar uma grande conservação da fenilalanina (F) que

sucede a glicina (Figura 10), sugerindo que a fenilalanina (F) possivelmente tenha um papel no direcionamento dessa curva. Não foram obtidas sequências com o motivo RGxxY, como observado por PEREIRA (2014)

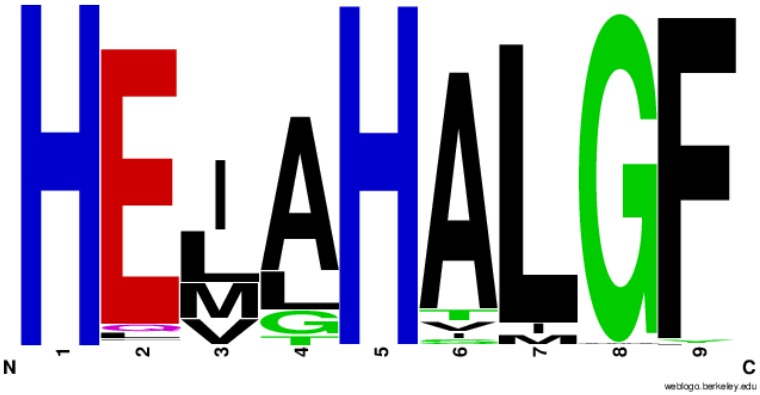

Figura 10: Representação gráfica (Logo) do início do motivo. Note a conservação das duas histidinas (H) nas posições 1 e 5, da glicina (G) e fenilalanina (F) nas posições 8 e 9 respectivamente e da alta freqüência do ácido glutâmico (E) na posição 2.

De acordo com SCHALANGENHAUF, ETGES e METCALF (1998) o cristal da *leishmanolysin* (GP63) de *L. major* (1LML) possui algumas alterações em seu domínio catalítico em relação à metaloproteases clássicas. Nas metaloproteases é possível observar uma molécula de água mantida próxima ao zinco através de duas ligações com os oxigênios do ácido glutâmico (E) do motivo HExxH, entretanto a densidade eletrônica do cristal 1LML indica a ausência dessa molécula de água nessa posição. Ao invés da molécula de água, a densidade eletrônica daquela região sugere a presença de uma glicina (G) de origem desconhecida que realiza duas ligações ao zinco em seu lugar, e mais três provenientes das três histidinas (H) do motivo, formando um total de cinco ligações ao zinco. Adicionalmente, o ácido glutâmico (E) do motivo HExxH adota uma conformação torcida para uma posição diferente das demais metaloproteases de zinco. Outra alteração descrita nessa metaloprotease é uma inserção de 62 aminoácidos entre a glicina (G) e a última histidina (H) no motivo, passando de HExxHxxGxxH para HExxHxxGx_{62}H. Isso sugere que o ácido glutâmico (E) não está mais envolvido no papel catalítico da proteína. Entretanto, um

experimento realizado por MACDONALD, MORRISON e MCMASTER (1995) com uma substituição do ácido glutâmico (E) do motivo por uma aspargina (N) em uma GP63 de *L. major* resultou na perda de atividade catalítica da enzima. É possível que a perda da atividade catalítica nesse experimento se dê ao fato de que a aspargina (N) é um aminoácido de carga positiva, enquanto o ácido glutâmico (E) é de carga negativa, o que poderia resultar em uma alteração na estrutura do motivo.

Um estudo feito por PEREIRA (2014) não obteve sucesso ao tentar produzir bactérias recombinantes *Escherichia coli* produtoras de GP63-1 (HExxH) e GP63-2 (HAxxH) de *T. rangeli*, pois essas proteínas comprovaram-se tóxicas e letais para a bactéria. Entretanto, nesse mesmo estudo foi comprovado que a atividade metaloprotease de *T. rangeli* é menor comparada a *T. cruzi* e *L. braziliensis.*

No modelo 1 produzido temos a alteração do ácido glutâmico (E265 no cristal 1LML) do motivo para alanina (A163), que é um aminoácido de carga neutra, o que não alterou a estrutura do motivo. Entretanto, a inserção entre a glicina (G169) e a última histidina (H227) é de apenas 57 aminoácidos, e a cadeia lateral desta histidina está voltada para o lado oposto ao do zinco (Figura 11), o que sugere uma alteração na funcionalidade da proteína, provavelmente reduzindo sua capacidade de se ligar ao zinco e prejudicando seu funcionamento. Entretanto, vale lembrar que o modelo gerado pelo I-Tasser não leva em consideração as ligações metálicas com o zinco, o que poderia acabar resultando em uma alteração no posicionamento da última histidina H227.

Já no modelo 2 é possível observar uma maior similaridade entre os motivos do modelo gerado ao cristal de 1LML, inclusive na conservação do ácido glutâmico (E168) (Figura 12). O tamanho da inserção entre a glicina (G174) e a última histidina (H236) do motivo é de 61 aminoácidos, um número muito mais próximo ao do cristal 1LML, e neste a histidina H236 encontra-se alinhada com a histidina H334 do cristal, sugerindo que não deve haver alteração na funcionalidade da proteína. Isso sugere que o tamanho da inserção entre a glicina (G) e a última histidina (H) do motivo em *T. rangeli* pode alterar o seu posicionamento em relação ao zinco e prejudicar a sua capacidade catalítica. Essa idéia é corroborada pela grande quantidade de sequências obtidas com a inserção perto de 62 aminoácidos, como visto na figura 9.

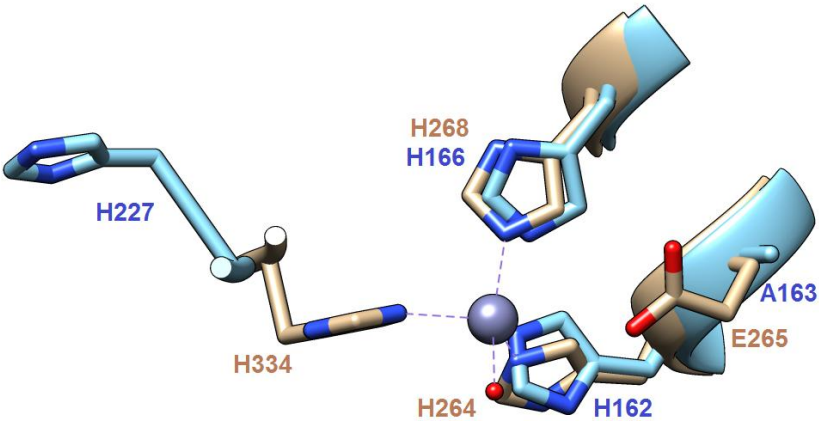

Figura 11: Representação do motivo do domínio catalítico do modelo 1 (em azul) com o cristal 1LML (em bege). A esfera cinza representa o zinco e a vermelha uma molécula de água. Note a histidina H227 do modelo 1 virada para o lado oposto ao zinco.

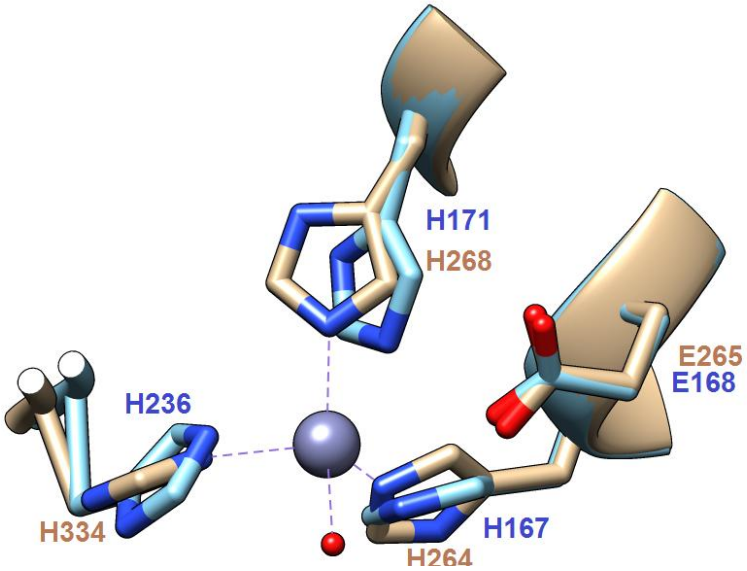

Figura 12: Representação do motivo do domínio catalítico do modelo 2 (em azul) com o cristal 1LML (em bege). A esfera cinza representa o zinco e a vermelha uma molécula de água. Note o alinhamento da histidina H236 do modelo 2 em relação a histidina H334 do cristal 1LML.

CALIXTO, BITAR, *et al.* (2013) realizaram uma modelagem comparativa de uma GP63 de *T. rangeli* com o motivo HExxH onde aparentemente o ácido glutâmico do motivo se liga com o zinco juntamente com as três histidinas do motivo completo. Entretanto, no modelo 2 é possível observar uma similaridade muito grande com a proteína 1LML, o que sugere que essa ligação não ocorra nessa proteína. Isso corrabora com a ideia de que uma substituição do ácido glutâmico (E) por um aminoácido neutro, como a alanina (A) visto no modelo 1, não prejudicaria a capacidade catalítica da proteína.

A carga da superfície do modelo 1 se mostrou ser mais positiva do que a carga vista no cristal 1LML (Figura 5). CALIXTO, BITAR, *et al.* (2013) realizou uma análise similar e sugeriu que a GP63 de *T. rangeli* deve possuir um substrato diferente de *T. cruzi* e *L. major* pois foi a única que apresentou uma carga de superfície mais positiva, enquanto as demais apresentaram uma carga predominantemente negativa.

As sequências utilizadas para a produção do modelo 1 e modelo 2 foram classificadas como pertencentes aos grupos de ortologia 4 e 2 respectivamente, que juntamente com o grupo 3 são compostas exclusivamente de sequências de *T. cruzi* e *T. rangeli*, o que sugere que as proteínas presentes nesses 3 grupos desempenham papéis importantes para essas espécies.

A conservação do ácido glutâmico (E) no motivo do grupo 2, assim como a conservação do tamanho da inserção, que varia pouco (61-62 aminoácidos) em relação a sequência do cristal de *leishmanolysin* 1LML (62 aminoácidos) sugerem que as sequências desse grupo possuem um motivo totalmente funcional, como observado no modelo tridimensional 2 (Figura 11). As sequências do grupo 3 apesar de terem o ácido glutâmico conservado no motivo possuem uma inserção consideravelmente maior (155 – 289 aminoácidos) após a glicina, o que muito provavelmente resulta em uma alteração drástica do posicionamento da última histidina do motivo no domínio catalítico, que pode acabar afetando a capacidade catalítica da proteína.

No grupo 4 todas as sequências possuem uma substituição do ácido glutâmico (E) por algum outro aminoácido. Apesar da subsituição do ácido glutâmico (E) pela alanina (A) na sequência GP63_B do modelo 1 não ter afetado a estrutura do domínio, não é possível afirmar que o mesmo ocorre com as sequências de *T. cruzi* desse grupo.

O grupo 1 apresentou a maior quantidade de sequências e maior diversidade de espécies, sugerindo a presença de duplicações gênicas e

genes parálogos nesse grupo. Além disso, foram encontradas algumas substituições do ácido glutâmico (E) no motivo HExxH por outros aminoácidos e uma variação significativa no tamanho da inserção. Devido a essa heterogeneidade entre as sequências desse grupo não é possível afirmar se há alterações significativas no domínio, visto que essa diferença no tamanho da inserção entre 58 e 62 resíduos pode ser significativa na funcionalidade da última histidina do motivo. Isso sugere que as sequências de GP63 pertencentes a esse grupo ortólogo não necessariamente desempenham sua atividade catalítica com a mesma eficácia. Essa observação é corroborada pela diversidade de espécies encontradas nesse grupo, visto que temos parasitas heteróxenos de invertebrados como *Leptomonas pyrrhocoris* e *Crithidia fasciculata*, juntamente com as demais espécies que são parasitas heteróxenos de mamíferos. Entretanto, a grande quantidade de sequências encontradas aqui sugere que esse é o grupo de ortologia com maior conservação entre as diferentes espécies.

6 CONCLUSÃO

●As alterações do domínio catalítico do modelo 1 de *T. rangeli* grupo B feito com a sequência de motivo HAxxH sugerem uma perda na função dessa proteína, principalmente pelo deslocamento da última histidina do motivo;

●O tamanho da inserção entre a glicina e a última histidina do motivo do domínio catalítico de GP63 em *T. rangeli* pode afetar o posicionamento da histidina, podendo prejudicar o funcionamento da proteína;

●O ácido glutâmico foi observado em maior número na sua posição em relação a outros aminoácidos, o que indica uma importância para o funcionamento dessa proteína. Entretanto, sua frequência foi menor que os demais aminoácidos característicos do motivo $HExxHxxGx_nH$, sugerindo uma menor importância deste aminoácido em relação aos demais, o que poderia ser um indício de que sua subsitituição não implica necessariamente na perda da função da proteína;

●Apesar da inserção presente no motivo em tripanosomatídeos, a glicina (G) e a fenilalanina (F) podem apresentar um papel importante para a organização estrutural da proteína;

●As proteínas GP63 com o motivo RGxxY provavelmente são proteínas exclusivas de *T. rangeli* e não foram encontradas nas buscas realizadas;

●Foram encontrados grupos de ortologia composto exclusivamente de sequências de *T. cruzi* e *T. rangeli*, sugerindo que essas são proteínas que podem desempenhar papéis importantes para essas duas espécies, enquanto o grupo 1 é composto por sequências de diversas espécies de tripanosomatídeos, sugerindo que as sequências encontradas dentro desse grupo são as mais conservadas entre as espécies.

7 REFERÊNCIAS

ALBERTS, B. et al. **Molecular Biology of the Cell**. 4ª ed. Nova Iorque: Garland Science, 2002.

ASLETT, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic Acids Res**, v. 38, 2010.

BANGS, J. D. et al. In vitro cytocidal effects on Trypanosoma brucei and inhibition of Leishmania major GP63 by peptidomimetic metalloprotease inhibitors. **Molecular & Biochemical Parasitology**, v. 114, 2001.

BERGERON, B. **Bioinformatics Computing**. [S.l.]: Prentice Hall PTR, 2002.

BERMAN, H. M. et al. The Protein Data Bank. **Nucleic Acid Res**, v. 28, 2000.

BERRIMAN, M. et al. The genome of the African trypanosome Trypanosoma brucei. **Science**, v. 309, n. 5733, p. 416-22, 2005.

CALIXTO, P. H. M. et al. Gene identification and comparative molecular modeling of a Trypanosoma rangeli major surface protease. **J Mol Model**, v. 19, p. 3053-3064, 2013.

CHANG, C. S.; CHANG, K. P. Monoclonal antibody affinity purification of a Leishmania membrane glycoprotein and its inhibition of leishmania-macrophage binding. **Proc Natl Acad Sci U S A**, v. 83, p. 100-104, 1986.

CROOKS, G. E. et al. WebLogo: A sequence logo generator. **Genome Research**, v. 14, p. 1188-1190, 2004.

CUEVAS, I. C.; CAZZULO, J. J.; SÁNCHEZ, D. O. Gp63 Homologues in Trypanosoma cruzi: Surface Antigens with Metalloprotease Activity and a Possible Role in Host Cell Infection. **Infection and Imunology**, v. 71, n. 10, 2003.

EL-SAYED, N. M. et al. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. **Science**, v. 309, n. 5733, p. 409-15, 2005.

EMANUELSSON, O. et al. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. **J. Mol. Biol.**, v. 300, p. 1005-1016, 2000.

FERREIRA, K. A. M. et al. Genome Survey Sequence Analysis and Identification of Homologs of Major Surface Protease (gp63) Genes in Trypanosoma rangeli. **Vector-Borne and zoonotic diseases**, v. 10, n. 9, 2010.

GRISARD, E. C. et al. Transcriptomic analyses of the avirulent protozoan parasite Trypanosoma rangeli. **Molecular & Biochemical Parasitology**, v. 174, 2010.

HALL, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. **Nucl. Acids. Symp. Ser.**, v. 41, p. 95-98, 1999.

IVENS, A. C. et al. The Genome of the Kinetoplastid Parasite, Leishmania major. **Science**, v. 309, n. 5733, p. 436-442, 2005.

KULKARNI, M. M. et al. Trypanosoma cruzi GP63 Proteins Undergo Stage-Specific Differential Posttranslational Modification and are Important for Host Cell Infection. **Infection and Immunity**, v. 77, n. 5, 2009.

LACOUNT, D. J. et al. Expression and Function of the Trypanosoma brucei Major Surface Protease (GP63) Genes, v. 278, n. 27, 2003.

LESK, A. M. **Introduction to Bioinformatics**. New York: Oxford University Press, 2002.

LI, L. et al. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. **Genome Res**, v. 13, p. 2178-2189, 2003.

MACDONALD, M. H.; MORRISON, C. J.; MCMASTER, W. R. Analysis of the active site and activation mechanism of the Leishmania surface metalloproteinase GP63. **Biochim Biophys Acta**, p. 199-207, 1995.

MCGWIRE, B.; CHANG, K. P. Genetic rescue of surface metalloproteinase (gp63)-deficiency in Leishmania amazonensis variants increases their infection of macrophages at the early phase. **Molecular and Biochemical Parasitology**, v. 66, 1994.

NEVES, D. P. et al. **Parasitologia Humana**. 12ª Edição. ed. São Paulo: Atheneu, 2011.

PEREIRA, T. K. S. **Caracterização de genes codificantes para metaloproteases (GP63) de Trypanosoma rangeli**. Universidade Federal de Santa Catarina. Florianópolis. 2014.

SANTOS, A. L. S.; BRANQUINHA, M. H.; D'VILLA-LEVY, C. M. The ubiquitous gp63-like metalloprotease from lower trypanosomatids: in the search for a function. **Anais da Academia Brasileira de Ciências**, v. 78, 2006.

SCHALANGENHAUF, E.; ETGES, R.; METCALF, P. The crystal structure of the Leishmania major surface proteinase leishmanolysin (gp63). **Structure**, v. 6, n. 8, 1998.

STOCO, P. H. et al. Genome of the Avirulent Human-Infective Trypanosome - Trypanosoma rangeli. **PLOS Neglected Tropical Diseases**, v. 8, n. 9, 2014.

VERLI, H. **Bioinformática:** da Biologia à Flexibilidade molecular. 1ª Edição. ed. São Paulo: Sociedade Brasileira de Bioquímica e Biologia Molecular, 2014.

YANG, J. et al. The I-TASSER Suit: Protein structure and function prediction. **Nature Methods**, v. 12, p. 7-8, 2015.

YAO, C. Major Surface Protease of Trypanosomatids: One Size Fits All? **Infection and Immunity**, v. 78, n. 1, p. 22-31, 2010.

YAO, C.; DONELSON, J. E.; WILSON, M. E. The major surface protease (MSP or GP63) of Leishmania sp. Biosynthesis, regulation of expression and function. **Molecular and Biochemical Parasitology 132**, p. 1-16, 2003.

APÊNDICE A – TABELA DOS GRUPOS DE ORTOLOGIA.

Legenda:

CFAC: *Crithidia fasciculata*; ENDO: *Endotrypanum monterogeii*; LAET: *Leishmania aethiopica*; LAMA: *Leishmania amazonensis*; LBRA: *Leishmania braziliensis*; LDON: *Leishmania donovani*; LENR: *Leishmania enrietti*; LGER: *Leishmania gerbilli*; LGUA: *Leishmania guyanensis*; LINF: *Leishmania infantum*; LMFR: *Leishmania major* Friedlin; LMLV: *Leishmania major* LV 3956; LMSD: *Leishmania major* SD75.1; LMEX: *Leishmania mexicana*; LPAN: *Leishmania panamensis*; LEIS: *Leishmania sp.;* LTAR: *Leishmania tarentolae;* LTRO: *Leishmania tropica*; LTUR: *Leishmania turanica*; LEPY: *Leptomonas pyrrhocoris*; LESE: *Leptomonas seymouri;* TBBT: *Trypanosoma brucei brucei* TREU 927; TBGD: *Trypanosoma brucei gambiense* DAL 972; TBRU: *Trypanosoma brucei rhodesiense*; TBBL: *Trypanosoma brucei brucei* Lister Strain 427; TCON: *Trypanosoma congolense*; TCRU: *Trypanosoma cruzi*; TCCL: *Trypanosoma cruzi* CL Brener Esmeraldo-like; TCNE: *Trypanosoma cruzi* CL Brener non Esmeraldo-like; TCMK: *Trypanosoma cruzi* Marinkellei; TCDM: *Trypanosoma cruzi* Dm28c; TCSY; *Trypanosoma cruzi* Sylvio x10/1; TEVN: *Trypanosoma evansi*; TGRA: *Trypanosoma grayi*; TRAN: *Trypanosoma rangeli* TVIV: *Trypanosoma vivax*

Grupo de ortologia	**Código da espécie | Access**	
Grupo 1	ENDO|EMOLV88_000010400	TCRU|EAN82216.1
	ENDO|EMOLV88_000010800	TCRU|EAN82430.1
	ENDO|EMOLV88_000011000	TCRU|EAN82680.1
	ENDO|EMOLV88_000011200	TCRU|EAN88454.1
	ENDO|EMOLV88_000011300	TCRU|EAO00007.1
	ENDO|EMOLV88_000012800	TCSY|EKG04985.1
	ENDO|EMOLV88_000012900	TCSY|EKG05166.1
	ENDO|EMOLV88_000019900	TCSY|EKG05355.1
	ENDO|EMOLV88_000020000	TEVN|TevSTIB805.11_01.8040
	ENDO|EMOLV88_000021800	TEVN|TevSTIB805.11_01.8060

	ENDO|EMOLV88_000060600	TEVN|TevSTIB805.8.1540
	ENDO|EMOLV88_280006900	TEVN|TevSTIB805.8.1550
	ENDO|EMOLV88_280007100	TGRA|KEG06574.1
	CFAS|CFAC1_140029100	TGRA|KEG06672.1
	CFAS|CFAC1_150013400	TGRA|KEG08395.1
	CFAS|CFAC1_200036800	TRAN|ESL05186.1
	CFAS|CFAC1_240049700	TRAN|ESL05237.1
	CFAS|CFAC1_260005400	TRAN|ESL05378.1
	CFAS|CFAC1_270047100	TRAN|ESL05601.1
	CFAS|CFAC1_270047200	TRAN|ESL07572.1
	CFAS|CFAC1_270047500	TVIV|CCC53338.1
	ENDO|EMOLV88_000010600	TVIV|CCC53361.1
	ENDO|EMOLV88_000010700	TVIV|CCC53434.1
	ENDO|EMOLV88_000010900	CFAS|CFAC1_040011500
	ENDO|EMOLV88_000019100	CFAS|CFAC1_040011600
	ENDO|EMOLV88_000020200	CFAS|CFAC1_040011700
	ENDO|EMOLV88_000037600	CFAS|CFAC1_040011800
	ENDO|EMOLV88_000038300	CFAS|CFAC1_040011900
	ENDO|EMOLV88_000057600	CFAS|CFAC1_040012000
	ENDO|EMOLV88_000058100	CFAS|Q06031.2
	ENDO|EMOLV88_280006700	ENDO|EMOLV88_000014700
	ENDO|EMOLV88_280006800	ENDO|EMOLV88_000019700
	ENDO|EMOLV88_280007200	ENDO|EMOLV88_000023600
	ENDO|EMOLV88_280015300	ENDO|EMOLV88_000024000
	LAET|LAEL147_000487400	ENDO|EMOLV88_000025200
	LDON|CBZ35465.1	ENDO|EMOLV88_000025400
	LEIS|LMARLEM2494_280012000	ENDO|EMOLV88_000025500
	LEPY|KPA78762.1	ENDO|EMOLV88_000031100
	LEPY|KPA78766.1	ENDO|EMOLV88_000032800
	LEPY|KPA78768.1	ENDO|EMOLV88_000042000
	LEPY|KPA80249.1	ENDO|EMOLV88_000048800
	LGER|LGELEM452_280011700	ENDO|EMOLV88_000048900

	LMFR|CAJ05048.1	ENDO|EMOLV88_000049300
	LMLV|LMJLV39_280011300	ENDO|EMOLV88_000049600
	LMSD|LMJSD75_280011300	ENDO|EMOLV88_000050300
	LTRO|LTRL590_280010900	ENDO|EMOLV88_000050500
	LTUR|LTULEM423_280011400	ENDO|EMOLV88_000050600
	TCCL|EAN91492.1	ENDO|EMOLV88_000050700
	TCCL|EAN93348.1	ENDO|EMOLV88_170015300
	TCDM|ESS55215.1	ENDO|EMOLV88_170018900
	TCMK|EKF30117.1	ENDO|EMOLV88_340021400
	TCNE|EAN82459.1	LBRA|CAM37248.1
	TCSY|EKG01091.1	LBRA|CAM37249.1
	TCSY|EKG04772.1	LBRA|CAM37252.1
	TVIV|CCC52269.1	LBRA|CAM37253.1
	TVIV|CCC52794.1	LBRA|CAM37254.2
	ENDO|EMOLV88_000024200	LBRA|CAM37255.1
	ENDO|EMOLV88_000024500	LBRA|CAM37256.1
	ENDO|EMOLV88_000024800	LBRA|CAM37257.1
	ENDO|EMOLV88_000049000	LBRA|CAM37259.1
	ENDO|EMOLV88_000049100	LBRA|CAM37260.1
	LAMA|AAA82695.1	LBRA|CAM37261.1
	LEPY|KEG09032.1	LBRA|CAM37262.1
	LEPY|KEG09800.1	LBRA|CAM37356.1
	LEPY|KEG14470.1	LBRA|CAM37358.1
	TBBL|Tb427.08.1610	LBRA|CAM37360.1
	TBBL|Tb427.08.1620	LBRA|CAM37362.1
	TBBL|Tb427.08.1630	LBRA|CAM37363.1
	TBBL|Tb427.08.1640	LBRA|CAM37364.1
	TBBT|AAZ12929.1	LBRA|CAM37365.2
	TBBT|AAZ12931.1	LBRA|CAM37366.1
	TBBT|AAZ12932.1	LBRA|CAM37367.1
	TBBT|CAJ17011.1	LBRA|CAM37368.1
	TBBT|CAJ17012.1	LBRA|CAM37370.1

	TBBT|CAJ17013.1	LBRA|LbrM.10.1580
	TBBT|EAN79736.1	LBRA|LBRM2903_000008900
	TBBT|EAN79738.1	LBRA|LBRM2903_000009800
	TBBT|EAN79739.1	LBRA|LBRM2903_100010500
	TBBT|Tb11.v5.0710	LBRA|LBRM2903_100010600
	TBBT|Tb927.11.7620	LDON|AAA29236.1
	TBBT|Tb927.11.7630	LDON|AAA29244.1
	TBBT|Tb927.11.7710	LDON|AAA53687.1
	TBGD|CBH13173.1	LDON|AAA53688.1
	TBRU|AAB61263.1	LDON|ACT31401.1
	TBRU|AAB61265.1	LDON|CAD42813.1
	TCCL|EAN81962.1	LDON|CBY93801.1
	TCCL|EAN82319.1	LDON|CBY93809.1
	TCCL|EAN82567.1	LDON|CBY93851.1
	TCCL|EAN83208.1	LEIS|LMARLEM2494_000024400
	TCCL|EAN84362.1	LEIS|LMARLEM2494_100009500
	TCCL|EAN84601.1	LENR|LENLEM3045_000005300
	TCCL|EAN84602.1	LENR|LENLEM3045_000009200
	TCCL|EAN87695.1	LENR|LENLEM3045_000011600
	TCCL|EAN87821.1	LENR|LENLEM3045_000013000
	TCCL|EAN88281.1	LENR|LENLEM3045_000020100
	TCCL|EAN88333.1	LENR|LENLEM3045_000022200
	TCCL|EAN88334.1	LENR|LENLEM3045_000026000
	TCCL|EAN90724.1	LENR|LENLEM3045_000028800
	TCCL|EAN94413.1	LENR|LENLEM3045_000035000
	TCCL|EAN95879.1	LENR|LENLEM3045_000036200
	TCCL|EAN98774.1	LENR|LENLEM3045_000038800
	TCCL|EAN99438.1	LENR|LENLEM3045_070005000
	TCDM|ESS55029.1	LENR|LENLEM3045_080005000
	TCDM|ESS64709.1	LENR|LENLEM3045_090005000
	TCMK|EKF26783.1	LENR|LENLEM3045_110005000
	TCMK|EKF27974.1	LENR|LENLEM3045_110005100

	TCMK|EKF28005.1	LENR|LENLEM3045_120005000
	TCMK|EKF28536.1	LENR|LENLEM3045_120017900
	TCMK|EKF29554.1	LENR|LENLEM3045_140021800
	TCMK|EKF31325.1	LENR|LENLEM3045_160005000
	TCMK|EKF36323.1	LENR|LENLEM3045_200024600
	TCMK|EKF36324.1	LENR|LENLEM3045_230027400
	TCMK|EKF38475.1	LENR|LENLEM3045_260035100
	TCNE|EAN81548.1	LENR|LENLEM3045_270005000
	TCNE|EAN81694.1	LENR|LENLEM3045_330043100
	TCNE|EAN82026.1	LEPY|KPA75955.1
	TCNE|EAN82048.1	LGUA|AAA29239.1
	TCNE|EAN82196.1	LGUA|AAA29240.1
	TCNE|EAN82508.1	LGUA|AAA29241.1
	TCNE|EAN82699.1	LINF|CAA69349.1
	TCNE|EAN82801.1	LINF|CAM66064.2
	TCNE|EAN83741.1	LINF|CAM66066.2
	TCNE|EAN83742.1	LINF|CAM66067.1
	TCNE|EAN85265.1	LINF|CAM66068.2
	TCNE|EAN85376.1	LINF|CBZ08391.1
	TCNE|EAN87126.1	LINF|CBZ08849.1
	TCNE|EAN88977.1	LMEX|CAA45733.1
	TCNE|EAN89350.1	LMFR|AAC39120.1
	TCNE|EAN90801.1	LMFR|ACL01096.2
	TCNE|EAN90802.1	LMFR|CAJ02586.1
	TCNE|EAN91158.1	LMFR|CAJ02588.1
	TCNE|EAN91657.1	LMFR|CAJ02589.1
	TCNE|EAN95336.1	LMFR|CAJ02591.1
	TCNE|EAN95957.1	LMLV|LMJLV39_100010200
	TCNE|EAN97548.1	LMLV|LMJLV39_100010300
	TCNE|EAN98248.1	LPAN|AIN96111.1
	TCNE|EAN99172.1	LTAR|LtaP10.0480
	TCON|CCC95375.1	LTAR|LtaP10.0650

	TCON|CCD15540.1	TCMK|EKF38779.1
	TCON|CCD15541.1	
Grupo 2	TCCL|EAN82338.1	TCNE|EAN88510.1
	TCCL|EAN86741.1	TCNE|EAN91658.1
	TCCL|EAN94015.1	TCNE|EAN91661.1
	TCCL|EAN96672.1	TCNE|EAN93807.1
	TCCL|EAN96694.1	TCNE|EAN95996.1
	TCCL|EAN96697.1	TCNE|EAN96630.1
	TCCL|EAN96699.1	TCNE|EAN96852.1
	TCCL|EAN98632.1	TCNE|EAN99558.1
	TCCL|EAN98641.1	TCNE|EAN99571.1
	TCCL|EAN99593.1	TCRU|EAN82147.1
	TCDM|ESS55555.1	TCRU|EAN83348.1
	TCDM|ESS55607.1	TCRU|EAN89556.1
	TCDM|ESS58247.1	TCRU|EAN92029.1
	TCDM|ESS60765.1	TCRU|EAN93302.1
	TCDM|ESS61287.1	TCRU|EAN93958.1
	TCDM|ESS61926.1	TCSY|EKF99467.1
	TCDM|ESS63082.1	TCSY|EKG02251.1
	TCNE|EAN82145.1	TRAN|ESL05225.1
	TCNE|EAN85955.1	TRAN|ESL05467.1
Grupo 3	TCCL|EAN89523.1	TCNE|EAN97742.1
	TCCL|EAN93698.1	TCNE|EAN97743.1
	TCMK|EKF37980.1	TCNE|EAN97752.1
	TCNE|EAN81389.1	TCSY|EKG05811.1
	TCNE|EAN84769.1	TCSY|EKG07009.1
	TCNE|EAN91909.1	TRAN|ESL05418.1
Grupo 4	TRAN|AGN32991.1	TCRU|EAN86911.1
	TCMK|EKF38831.1	TCCL|EAN96526.1
	TCNE|EAN88927.1	TRAN|GP63BMODEL
Grupo 5	LBRA|CAM42646.1	LPAN|LPAL13_310025300
	LBRA|LBRM2903_310032100	LEIS|LMARLEM2494_000013700

	LTAR|LtaP31.2430	LESE|KPI87511.1
	LTRO|LTRL590_000018300	LEPY|KPA78759.1
	LTUR|LTULEM423_310026600	TBBT|EAN79824.1
	LDON|CBZ36604.1	TBBL|Tb427tmp.12.0006
	LINF|CAM70552.1	LARA|LARLEM1108_310028100
	LGER|LGELEM452_310026600	TCON|CCC95446.1
	LMFR|CAJ08388.1	ENDO|EMOLV88_310024800
	LMSD|LMJSD75_310028800	LAET|LAEL147_000607400
	LMLV|LMJLV39_310028700	TBGD|CBH17856.1
	LMEX|CBZ29256.1	CFAS|CFAC1_270047600
	LENR|LENLEM3045_310015500	
Grupo 6	TGRA|KEG05606.1	TGRA|KEG10236.1
	TGRA|KEG12605.1	TGRA|KEG13227.1
	TGRA|KEG07354.1	TGRA|KEG09903.1
	TGRA|KEG05971.1	LEPY|KEG09903.1
	TGRA|KEG06833.1	LEPY|KEG13227.1
	TGRA|KEG07047.1	LEPY|KEG10236.1
	TGRA|KEG07190.1	LEPY|KEG10238.1
	TGRA|KEG06676.1	LEPY|KEG12605.1
Grupo 7	TBBT|EAN79707.1	TGRA|KEG12488.1
	TBBT|Tb927.11.7410	TVIV|CCC53317.1
	TCCL|EAN91476.1	TCON|CCC95354.1
	TCNE|EAN90139.1	TBBL|Tb427tmp.02.5310
	TCDM|ESS64459.1	TBGD|CBH17729.1
	TCSY|EKG07113.1	LEPY|KEG12488.1
Grupo 8	TCRU|AAL86597.1	TCRU|EAN86625.1
	TCRU|EAN86645.1	TCRU|EAN82728.1
	TCRU|EAN81991.1	TCMK|EKF29576.1
	TCRU|EAN98328.1	TCNE|EAN93610.1
	TCRU|EAN87763.1	TCDM|ESS55066.1
Grupo 9	TVIV|CCD20103.1	TCON|CCC93446.1
	TVIV|CCD20285.1	TEVN|TevSTIB805.10.2590

	TVIV|CCD20174.1	TBBT|EAN77667.1
	TVIV|CCD18177.1	TBGD|CBH15211.1
Grupo 10	TCMK|EKF31021.1	TGRA|KEG07421.1
	TCNE|EAN98730.1	TCCL|EAN88295.1
	TCSY|EKF99824.1	TCDM|ESS67097.1
Grupo 11	TGRA|KEG11903.1	LEPY|KEG11903.1
	TGRA|Tgr.217.1100	
Grupo 12	LEPY|KPA86218.1	CFAS|CFAC1_300073300
	LESE|KPI90271.1	
Grupo 13	LEPY|KEG10703.1	TGRA|KEG10703.1
Grupo 14	LEPY|KEG10866.1	TGRA|KEG10866.1
Grupo 15	LEPY|KEG10867.1	TGRA|KEG10867.1
Grupo 16	LEPY|KEG11581.1	TGRA|KEG11581.1
Grupo 17	LEPY|KEG15352.1	TGRA|KEG15352.1