MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado em Biologia Computacional e Sistemas

# DE SUPERGRUPOS A SUPERFAMÍLIAS, UM ESTUDO DE HOMOLOGIA EM PROTOZOÁRIOS.

DARUECK ACÁCIO CAMPOS

Rio de Janeiro
Julho de 2018

Ministério da Saúde

**FIOCRUZ**
**Fundação Oswaldo Cruz**

**INSTITUTO OSWALDO CRUZ**
**Programa de Pós-Graduação em Biologia Computacional e Sistemas**

*DARUECK ACÁCIO CAMPOS*

# De Supergrupos a Superfamílias, um estudo de homologia em protozoários.

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Biologia Computacional e Sistemas

**Orientador:** Dr. Alberto Martín Rivera Dávila

**RIO DE JANEIRO**
Julho de 2018

Campos, Darueck Acácio.

De supergrupos a superfamílias: um estudo de homologia em protozoários. / Darueck Acácio Campos. - Rio de janeiro, 2018.
132 f.; il.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2018.

Orientador: Alberto Martín Rivera Dávila.

Bibliografia: f. 77-95

1. Homologia. 2. Genômica. 3. Protozoários. 4. Métodos Computacionais. 5. Reconciliação. I. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/ICICT com os dados fornecidos pelo(a) autor(a).

Ministério da Saúde

**FIOCRUZ**
**Fundação Oswaldo Cruz**

## INSTITUTO OSWALDO CRUZ

**Programa de Pós-Graduação em Biologia Computacional e Sistemas**

***AUTOR: DARUECK ACÁCIO CAMPOS***

**DE SUPERGRUPOS A SUPERFAMÍLIAS: UM ESTUDO DE HOMOLOGIA EM PROTOZOÁRIOS.**

**ORIENTADOR:** **Dr. Alberto Martín Rivera Dávila**

**Aprovada em: 09/07/2018**

**EXAMINADORES:**

**Prof. Dr.** Milton Ozório Moraes **- Presidente**
**Prof. Dr.** Sérgio Manuel Serra da Cruz (UFRRJ/RJ)
**Prof. Dr.** Roney Santos Coimbra (FIOCRUZ/MG)
**Prof. Dr.** Maria Luiza Machado Campos (UFRJ)
**Prof. Dr.** Fabrício Alves Barbosa da Silva (PROCC/FIOCRUZ/RJ)

Rio de Janeiro, 9 de julho de 2018

Aos amores da minha vida

## AGRADECIMENTOS

**Agradeço à minha família, minha linda esposa Denila, meus filhos Thor, Lucas e a(o)(s) que ainda virá(virão), meus pais Aroldo e Eliane, minha avó Raimundinha, minhas irmãs, sobrinhas e sobrinho, sogros e cunhados, pelo enorme e irrestrito amor, apoio e compreensão e por segurarem "a barra" nos períodos que não pude estar presente.**

**Ao Programa de Pós-Graduação em Biologia Computacional e Sistemas, extensivo a todos os funcionários que passaram pela Secretaria Acadêmica durante esses quase cinco anos, pela disponibilidade e paciência sempre que solicitada.**

**Ao Instituto Federal De Educação, Ciência E Tecnologia Do Acre pelo apoio e oportunidade. Aos colegas e amigos que fiz no meu trabalho, e que me apoiaram nessa empreitada.**

**À cada uma das pessoas que eu tenho a honra de chamar de amigo. Parte da família escolhida. Parte do que eu sou e passageiro do mesmo vagão que o meu nesse trem da vida.**

**Aos meus amigos do Acre, que sei quer torceram muito por mim, assim como torço por eles.**

**Ao Dr. Alberto Dávila pela orientação, apoio, amizade e companheirismo ao me receber no Rio de janeiro como seu orientando.**

**Aos amigos que fiz durante esse doutorado e que contribuíram de forma fundamental para essa tese: Rodrigo, Bernardo, Rafael, Elisa, Roney, Fabrício, Mota, Diogo, Nelson, Ricardo, Antônio e Frazão.**

**Aos amigos que fiz fora do doutorado e que me receberam tão bem em seus lares em momentos importantes para minha estrutura mental, física e emocional: Marli, Alberto, Patrícia, Geraldo (best beer ever!), Juninho, Cida e seus filhos.**

**Obrigado!!!**

## EPÍGRAFE

*No fundo, o que importa é bem viver o seu dia, porque é com dias que se faz uma vida, irmão!*

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

# INSTITUTO OSWALDO CRUZ

**DE SUPERGRUPOS A SUPERFAMÍLIAS, UM ESTUDO DE HOMOLOGIA EM PROTOZOÁRIOS.**

## RESUMO

**TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS**

**Darueck Acácio Campos**

Protozoários patogênicos causam doenças importantes em países tropicais, como malária, doença do sono, doença de Chagas, leishmaniose, amebíase e giardíase, que em conjunto ameaçam milhões de pessoas em todo o mundo. Além disso, a maioria das doenças parasitárias causadas por protozoários são zoonóticas. Compreender a biologia desses organismos é crucial para combater as doenças que eles causam e estudos de genômica comparativa podem ser úteis para entender a relação evolutiva entre eles. Usando inferência de genômica comparativa e homologia, o presente estudo contemplou três espécies de protozoários de diferentes filos: *Cryptosporidium muris* (Apicomplexa), *Entamoeba invadens* (Amoebozoa) e *Trypanosoma grayi* (Euglenozoa), escolhidos por serem patógenos ainda pouco estudados e pela distância genética entre eles. A tese pode ser dividida em 3 partes. Numa primeira parte os programas de inferência de homologia OMA e OrthoMCL foram utilizados para inferir genes homólogos e seus resultados foram comparados e separados em 3 categorias de acordo com o nível de concordância entre eles, com ênfase na identificação de grupos homólogos com maior distância evolutiva e na identificação de multidomínios CDD (Conserved Domain Database) e Pfam-A (Pfam protein families database). Na segunda parte, propomos uma nova abordagem para a identificação de homólogos, com base na definição de "Supergrupos" homólogos, formados pela reconciliação dos resultados de ambos os programas; usando como critério para inferência a interseção de proteínas e para sua validação critérios de alta estringência, onde todas as proteínas (100%) do Supergrupo devem (a) ter o mesmo domínio conservado (CDD) identificado ou (b) pertencerem à mesma família de proteínas (Pfam-A). Na terceira e última parte, foi feita uma busca por genes homólogos distantes entre os mesmos protozoários de diferentes filos utilizados no primeiro e no segundo estudo utilizando comparação entre perfis do Modelo Oculto de Markov (pHMM - pHMM) com o programa de inferência de homologia COMA, visando a identificação de superfamílias de proteínas utilizando a base de dados de famílias e superfamílias de proteínas SUPERFAMILY. Nossos resultados mostraram que foi possível inferir novos grupos de proteínas homólogas utilizando as abordagens de reconciliação (Supergrupos homólogos) e de comparação pHMM – pHMM (Novos grupos homólogos distantes).

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

# INSTITUTO OSWALDO CRUZ

**FROM SUPERGROUPS TO SUPERFAMILIES, A STUDY OF PROTOZOAN HOMOLOGY.**

## ABSTRACT

**PHD THESIS IN COMPUTATIONAL BIOLOGY AND SYSTEMS**

**Darueck Acácio Campos**

Pathogenic protozoa cause major diseases in tropical countries, such as malaria, sleeping sickness, Chagas disease, leishmaniasis, amebiasis and giardiasis, which together threaten millions of people worldwide. In addition, most parasitic diseases caused by protozoa are zoonotic. Understanding the biology of these organisms is crucial in combating the diseases they cause, and studies of comparative genomics may be helpful in understanding the evolutionary relationship between them. Using comparative genomic inference and homology, the present study aimed at three protozoan species of different phyla: *Cryptosporidium muris* (Apicomplexa), *Entamoeba invadens* (Amoebozoa) and *Trypanosoma grayi* (Euglenozoa), chosen as pathogens that have not yet been studied and the genetic distance between them. The thesis can be divided into 3 parts. In a first part the inference programs OMA and OrthoMCL were used to infer homologous genes and their results were compared and separated into 3 categories according to the level of agreement between them, with emphasis on the identification of homologous groups with greater evolutionary distance and in the CDD (Conserved Domain Database) and Pfam-A (Pfam protein families database) multidomain identification. In the second part, we propose a new approach for the identification of homologues, based on the definition of homologous "Supergroups", formed by the reconciliation of the results of both programs; Using as criterion for inference the intersection of proteins and for their validation criteria of high stringency, where all proteins (100%) of the Supergroup must (a) have the same conserved domain (CDD) identified or (b) belong to the same protein family (Pfam-A). In the third and final part, a search was made for distant homologous genes between the same protozoa of different phyla used in the first and second studies using a comparison of profiles of the Occult Markov Model (pHMM - pHMM), with the inference program COMA, aiming at the identification of superfamilies of proteins using the database of families and superfamilies of proteins SUPERFAMILY. Our results showed that it was possible to infer new groups of homologous proteins using the reconciliation (Supergroup homologous) and the pHMM - pHMM (New distant homologous groups) approaches.

# ÍNDICE

**RESUMO** **IX**

**ABSTRACT** **X**

**1 INTRODUÇÃO** **1**

**1.1 Protozoários** ..... **1**

1.1.1 Protozoários utilizados nesse estudo ..... 3

**1.2 Homologia** ..... **5**

1.2.1 Métodos para a inferência de homologia ..... 7

1.2.2 Bases de dados de homólogos ..... 11

1.2.3 Outras bases de dados biológicos ..... 13

**1.3 Organização da tese** ..... **15**

**2 OBJETIVOS** **17**

**2.1 Objetivo geral** ..... **17**

**2.2 Objetivos específicos** ..... **17**

2.2.1 Identificar homólogos distantes entre as 3 espécies de protozoários utilizando dois programas para inferência de homologia com diferentes metodologias, além de realizar comparações entre seus resultados. ..... 17

2.2.2 Desenvolver um método para inferência de homologia baseada na reconciliação de grupos homólogos inferidos por duas abordagens distintas ..... 17

2.2.3 Validar os homólogos distantes inferidos através da comparação pHMM - pHMM com a classificação de superfamílias na base de dados Superfamily, verificar a quais superfamílias de proteínas estes poderiam estar associados e analisar funcionalmente os novos grupos homólogos distantes formados pela junção dos melhores hits recíprocos pHMM – pHMM. ..... 17

**3 MATERIAL E MÉTODOS** **18**

**3.1 Inferência de grupos homólogos em protozoários evolutivamente distantes com os programas OrthoMCL e OMA** ..... **18**

3.1.1 Caracterização e processo de criação de solução do estudo ..... 18
3.1.2 Genomas completos utilizados ..... 20
3.1.3 Identificação de homólogos através dos programas OMA e OrthoMCL ..... 20
3.1.4 Análise dos resultados do OMA e OrthoMCL ..... 21
3.1.5 Inferência de distância evolutiva do OMA e OrthoMCL ..... 21
3.1.6 Identificação de domínios conservados (CDD) ..... 21
3.1.7 Identificação de famílias de proteínas (Pfam-A) ..... 22
**3.2 Inferência de homologia baseada em uma abordagem de reconciliação para a genômica comparativa de protozoários. ..... 23**
3.2.1 Caracterização e processo de criação de solução do estudo ..... 23
3.2.2 Algoritmo de reconciliação – Inferência de Supergrupos ..... 24
3.2.3 Validação por CDD ou Pfam-A ..... 25
3.2.4 Filogenia ..... 25
3.2.5 Comparação com bases de dados SUPERFAMILY e Pfam Clans ..... 25
3.2.6 Análise da conservação das sequências ..... 25
3.2.7 Inferência da distância evolutiva nos Supergrupos ..... 25
3.2.8 Categorização funcional ..... 26
3.2.9 Vias metabólicas do KEGG ..... 26
**3.3 Inferência e análise de homólogos distantes em Protozoa por meio da comparação pHMM - pHMM (perfis de Modelo Oculto de Markov) visando a identificação de superfamílias ..... 27**
3.3.1 Caracterização e processo de criação de solução do estudo ..... 27
3.3.2 Comparação pHMM - pHMM com o COMA ..... 27
3.3.3 Validação dos homólogos distantes inferidos pelo COMA ..... 28
3.3.4 Inferência da distância evolutiva para os homólogos distantes inferidos via pHMM - pHMM ..... 29

**4 RESULTADOS 30**
**4.1 Inferência de grupos homólogos em protozoários evolutivamente distantes com os programas OrthoMCL e OMA ..... 30**

4.1.1 Inferência de homologia .....30
4.1.2 Análise dos resultados do OMA e OrthoMCL.....32
4.1.3 Inferência de distância evolutiva para o OrthoMCL e OMA.....33
4.1.4 Identificação de domínios conservados (CDD) .....35
4.1.5 Identificação de famílias de proteínas (Pfam-A).....35
**4.2 Inferência de homologia baseada em uma abordagem de reconciliação para a genômica comparativa de protozoários .....37**
4.2.1 Resultados da validação dos homólogos da Categoria II: "Divergentes com as interseções" – Inferência de Supergrupos .....37
4.2.2 Identificação de domínios conservados (CDD) nos Supergrupos .....40
4.2.3 Identificação de famílias de proteínas (Pfam-A) nos Supergrupos .....40
4.2.4 Comparação com as bases de dados SUPERFAMILY e Pfam Clans .....41
4.2.5 Análise da conservação das sequências dos Supergrupos .....42
4.2.6 Filogenia .....42
4.2.7 Inferência da distância evolutiva para os Supergrupos .....42
4.2.8 Categorização funcional dos Supergrupos.....44
4.2.9 Vias Metabólicas do KEGG dos Supergrupos.....45
**4.3 Inferência e análise de homólogos distantes em Protozoa por meio da comparação pHMM - pHMM (perfis de Modelo Oculto de Markov) visando a identificação de superfamílias.....46**
4.3.1 Comparação pHMM - pHMM com o COMA .....46
4.3.2 Validação dos novos homólogos distantes inferidos pelo COMA.....47
4.3.3 Criação de grafos para análises adicionais .....51
4.3.4 Inferência de distância evolutiva nos novos homólogos distantes .....57

**5 DISCUSSÃO 58**
**5.1 Inferência de Grupos Homólogos em Protozoários Evolutivamente Distantes Com os Programas OrthoMCL e OMA.....58**
5.1.1 Inferência de homologia .....58

5.1.2 Inferência de distância evolutiva.....59
5.1.3 Identificação de domínios conservados (CDD) .....59
5.1.4 Identificação de famílias de proteínas (Pfam-A).....60
**5.2 Inferência de homologia baseada em uma abordagem de reconciliação para a genômica comparativa de protozoários.....62**
5.2.1 Algoritmo de Reconciliação – Inferência de Supergrupos.....62
5.2.2 Análises funcionais.....64
**5.3 Inferência e análise de homólogos distantes em Protozoa por meio da comparação pHMM - pHMM (perfis de Modelo Oculto de Markov) visando a identificação de superfamílias.....67**
5.3.1 Comparação pHMM - pHMM com o COMA .....67
5.3.2 Validação dos homólogos distantes inferidos pelo COMA.....68
5.3.3 Análises funcionais das superfamílias a serem associadas aos grupos homólogos OrthoMCL/OMA alvos de reanotação funcional .....70
5.3.4 Análises funcionais das superfamílias mais frequentemente identificadas nos novos grupos homólogos distantes (core) .....72

**6 CONCLUSÕES 75**

**7 PERSPECTIVAS 77**

**8 REFERÊNCIAS BIBLIOGRÁFICAS 78**

**9 APÊNDICES E/OU ANEXOS 97**

**ÍNDICE DE FIGURAS**

**Figura 1-1: Exemplos de diferentes protozoários 1**

**Figura 1-2: Relações de homologia, paralogia e ortologia utilizando o gene da hemoglobina. 6**

**Figura 1-3: Representação de superfamílias usando múltiplos HMM. 10**

**Figura 3-1: Árvore filogenética dos maiores grupos evolucionários de eucariotos. 19**

**Figura 3-2: Fluxograma que detalha o processo de inferência de grupos homólogos do OrthoMCL e do OMA e realização de análises. 19**

**Figura 3-3: Fluxograma que detalha o processo de inferência dos Supergrupos e suas análises funcionais. 24**

**Figura 3-4: O algoritmo de reconciliação. 25**

**Figura 3-5: Fluxograma que detalha o processo de inferência de homólogos distantes e sua validação via SUPERFAMILY. 28**

**Figura 4-1: Porcentagem do genoma de cada espécie contribuindo para os grupos homólogos. 31**

**Figura 4-2: Diagrama de caixa apresentando os valores das distâncias evolutivas dentro dos grupos OMA e dos grupos do OrthoMCL que possuem somente ortólogos. 34**

**Figura 4-3: Domínios conservados (CDD) identificados em cada genoma. 35**

**Figura 4-4: Famílias de proteínas (Pfam-A) identificadas em cada genoma. 36**

**Figura 4-5: Validação dos Supergrupos por domínio conservado (CDD) e família de proteínas (Pfam-A). 38**

**Figura 4-6: Nova distribuição dos grupos homólogos inferidos neste estudo. 39**

**Figura 4-7: Distribuição das 116 proteínas presentes nos 22 Supergrupos homólogos validados por CDD. 40**

**Figura 4-8: Distribuição das 315 proteínas presentes nos 53 Supergrupos homólogos validados por Pfam-A. 41**

**Figura 4-9: Diagrama de caixa representando as distâncias evolutivas dos Supergrupos homólogos e dos grupos OrthoMCL e OMA originais. 43**

**Figura 4-10: Categorias funcionais inferidas para as proteínas dos 59 Supergrupos. 45**

**Figura 4-11: Resultado da validação dos grupos homólogos distantes prováveis identificados entre melhores hits recíprocos entre os grupos**

**parálogos inferidos pelo OrthoMCL, utilizando a base de dados SUPERFAMILY. 48**

**Figura 4-12: Resultado da validação dos grupos homólogos distantes prováveis identificados entre melhores hits recíprocos entre os grupos ortólogos inferidos pelo OrthoMCL, utilizando a base de dados SUPERFAMILY. 49**

**Figura 4-13: Resultado da validação dos grupos homólogos distantes prováveis identificados entre melhores hits recíprocos pHMM – pHMM entre os grupos ortólogos inferidos pelo OMA, utilizando a base de dados SUPERFAMILY. 50**

**Figura 4-14: Grafo representando as relações (arestas) entre os 114 grupos do OrthoMCL com ortólogos dos 3 organismos (nós) que tiveram melhor hit recíproco pHMM – pHMM. A figura mostra, quando possível, as relações entre os 3 melhores hits pHMM – pHMM e a superfamília identificada (código numérico e cor) em cada grupo. 53**

**Figura 4-15: Grafo representando as relações (arestas) entre os 92 grupos do OMA com ortólogos dos 3 organismos (nós) que tiveram melhor hit recíproco pHMM – pHMM. A figura mostra, quando possível, as relações entre os 3 melhores hits pHMM – pHMM e a superfamília identificada (código numérico e cor) em cada grupo. 54**

**Figura 4-16: Superfamílias identificadas nos novos grupos homólogos distantes OrthoMCL/COMA e OMA/COMA com ortólogos de *C. muris, E. invadens* e *T. grayi*. 56**

**Figura 4-17: Diagrama de caixa representando as distâncias evolutivas dos novos grupos homólogos distantes OrthoMCL/COMA e OMA/COMA e dos grupos OrthoMCL e OMA originais: 57**

# LISTA DE TABELAS

**Tabela 1: Genomas completos utilizados nos experimentos da tese. ..... 20**
**Tabela 2: Grupos homólogos inferidos pelo OrthoMCL e o OMA. ..... 32**
**Tabela 3: Distribuição dos grupos homólogos em 3 categorias de acordo com o nível de concordância entre as ferramentas de inferência de homologia. .... 32**
**Tabela 4: Supergrupos escolhidos para serem usados como caso de estudo por pertencerem à via "metabolismo". ..... 45**
**Tabela 5: Grupos OrthoMCL/OMA alvos de reanotação funcional..... 51**
**Tabela 6: Comparação entre alguns métodos de inferência de homologia distante. ..... 64**

# LISTA DE SIGLAS E ABREVIATURAS

BLAST – do inglês *Basic Local Alignment Search Tool*

CDD – do inglês *Conserved Domain Database*

Pfam – do inglês *Protein Family*

COG – do Inglês *Cluster of Orthologous Groups*

DNA - Ácido desoxirribonucléico

*E-value* – valor de probabilidade de um resultado ter sido obtido ao acaso, do inglês *Expectation value*

P-value - nível descritivo ou probabilidade de significância

KEGG – do inglês *Kyoto Encyclopedia of Genes and Genomes*

RefSeq – do inglês *Reference Sequence*

NCBI – do inglês *National Center of Biotechnology Information*

eggNOG – do inglês evolutionary genealogy of genes: Non-supervised Orthologous Groups

HMM - inglês Hidden Markov Models ou “modelos ocultos de Markov”

# 1 INTRODUÇÃO

## 1.1 Protozoários

Os protozoários são eucariotos unicelulares de vida livre que possuem uma grande variedade e complexidade estrutural, se adaptam a diversas condições ambientais (Ruppert, Fox & Barnes, 2004) e podem rapidamente alterar sua forma e englobar outras células por fagocitose (Alberts et al., 2014). Sua anatomia é frequentemente heterogênea, com estruturas como fotorreceptores, cerdas sensoriais, cílios com movimentação sinuosa, parte de boca, apêndices similares a pernas, feixes contráteis similares a músculos e dardos urticantes (Alberts et al., 2014) (Figura 1-1).

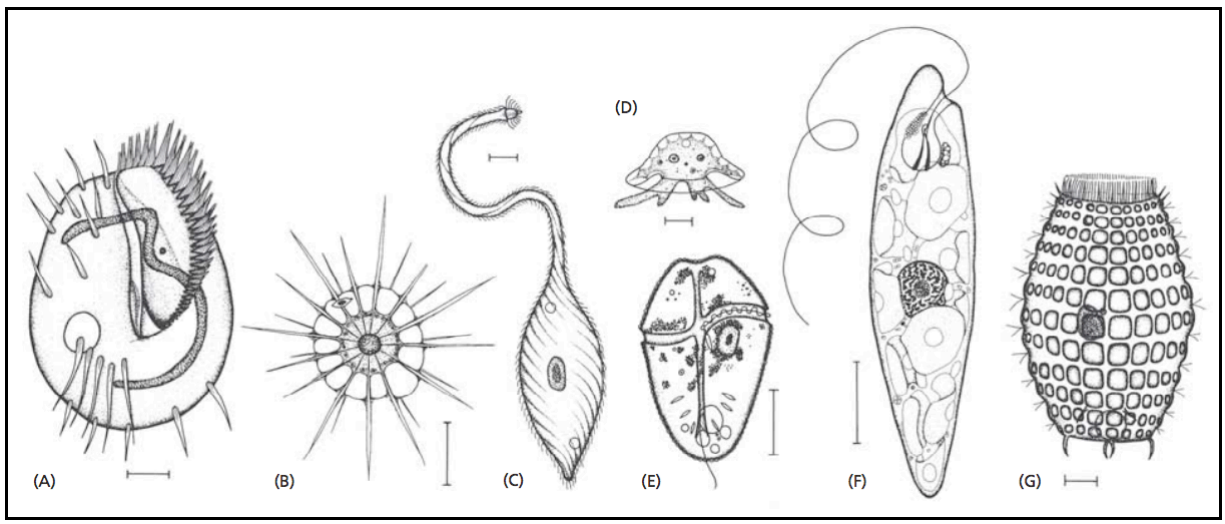



**Figura 1-1: Exemplos de diferentes protozoários**
As ilustrações foram feitas em diferentes escalas, mas, em cada caso, a barra de escala representa 10 µm. Os organismos em (A), (C) e (G) são ciliados; (B) é um heliozoário; (D) é uma ameba; (E) é um dinoflagelado; e (F) é uma euglena. (Fonte: Sleigh, 1973; Alberts et al., 2014)

O ciclo de vida dos protozoários parasitas é complexo, necessitando frequentemente de mais de um hospedeiro (Alberts et al., 2014). Acredita-se que existam aproximadamente de 200.000 espécies nomeadas de protozoários, das quais cerca de 10.000 seriam parasitas (Gransden, 1999). Além disso, os protozoários são capazes de se multiplicar em humanos, o que pode contribuir para sua sobrevivência. As infecções parasitárias por protozoários constituem uma das causas mais importantes de mortalidade e morbidade em seres humanos em grande

parte do mundo (CDC, 2016). Segundo a Organização Mundial da Saúde (OMS), dentre as principais doenças causadas por protozoários, temos a malária, com estimativa de 216 milhões de casos e 445.000 mortes em 2016, a leishmaniose com 700.000 – 1 milhão de casos e 20.000 – 30.000 mortes anuais (e mais de 1 bilhão de pessoas vivendo em áreas endêmicas), e a doença de Chagas com estimativa de 6 a 7 milhões de pessoas infectadas no mundo (WHO, 2018a,b). Estas doenças e outras (por exemplo: amebíase, giardíase, toxoplasmose e tricomoníase) apresentam um aumento de casos de refratividade ao tratamento principal. O fracasso do tratamento tem, potencialmente, uma origem multifatorial, sendo que uma das principais preocupações é a resistência a drogas (Burri & Keiser, 2001; Sobel, Nagappan, & Nyirjesy, 1999). Os parasitas protozoários que são infecciosos em humanos representam uma ameaça significativa à saúde e causam mais de um milhão de mortes por ano (Lozano et al., 2012).

Além disso, a maioria das doenças parasitárias clássicas causadas por protozoários são zoonóticas (Krauss et al, 2003). Existem ainda outras enfermidades também associadas a protozoários que apresentam impacto relevante não somente na saúde humana, mas também em outras espécies associadas ao ser humano, pois atingem animais utilizados para consumo ou convívio, como a babesiose, teileriose e a criptosporidíase (Brayton et al., 2007; Elmore et al., 2010; Pain et al., 2005; Widmer et al., 2009). Compreender a biologia desses organismos é crucial para combater as doenças que eles causam e estudos de genômica comparativa podem ser úteis para entender a relação evolutiva entre eles. O estudo da evolução do parasitismo é um problema central na biologia evolutiva (Jackson et al., 2016). Segundo Koonin & Gabaldón (2013), um dos pontos principais para a compreensão da biologia de organismos é ter seu genoma anotado.

A anotação funcional de um genoma consiste na identificação de suas regiões funcionais ou de relevância biológica e é possível realizar a transferência de anotação via técnicas computacionais de análise de similaridade de sequência e inferência de homologia, sendo este um processo essencial para a biologia moderna (Koonin & Gabaldón, 2013 ).

Nos últimos anos, como resultado do trabalho de várias equipes de pesquisadores, os genomas de 80 espécies de protozoários foram totalmente sequenciados (RefSeq, 2018), proporcionando uma boa base de dados para transferência de anotação. Porém, o processo de anotação de genomas pode

produzir um número considerável de erros, e o desenvolvimento de técnicas mais sensíveis se faz necessário (Koonin & Galperin, 2003).

### 1.1.1 Protozoários utilizados nesse estudo

Com o objetivo de realizar testes sobre a inferência de grupos homólogos em espécies de protozoários distantes, foram feitos estudos utilizando espécies de protozoários pertencentes a diferentes filos: *Cryptosporidium muris* (Apicomplexa), *Entamoeba invadens* (Amoebozoa) e *Trypanosoma grayi* (Euglenozoa) (Figura 3-1).

#### *1.1.1.1 Trypanosoma grayi*

*O Trypanosoma grayi* é um parasito extracelular da corrente sanguínea de crocodilos africanos (*Crocodylus niloticus*), causando tripanossomíase, com infecção adquirida via moscas tsé-tsé (*Glossina morsitans*) infectadas (Fermino et al., 2013). Abordagens filogenômicas realizados por Kelly (Kelly et al., 2014), demonstraram que o *Trypanosoma grayi* está mais proximamente relacionado ao *Trypanosoma cruzi* (transmissor da doença de Chagas em humanos) do que aos tripanosomas africanos *Trypanosoma brucei* (transmissor tripanossomíase africana em humanos), *Trypanosoma congolense* e *Trypanosoma vivax*, apesar do fato de *T. grayi* e os tripanosomas africanos também serem transmitidos por moscas tsé-tsé.

Uma vez que foram publicadas as sequências completas dos genomas de vários protozoários, incluindo *Leishmania major* (Ivens et al., 2005), *Trypanosoma cruzi* (El-Sayed et al., 2005), *Trypanosoma Brucei* (Berriman et ai., 2005) e, mais recentemente, *Trypanosoma grayi* (Kelly et al., 2014), pertencentes ao clado dos *Tripanosomatídeos*, esses dados genômicos puderam ajudar a aumentar a nossa compreensão sobre a evolução dessas espécies, sua estratégia primária de evasão imune, e também a evolução de suas moléculas de superfície celular que representam a interface hospedeiro-parasita (Jackson et al., 2012; Jackson et al., 2013). Estudos de genômica comparativa utilizando genoma completo de cepas diferentes de *Trypanosoma brucei* permitiram a identificação de associações significativas de hospedagem e localização geográfica. Foi detectada uma forte seleção de purificação em regiões genômicas associadas à estrutura do citoesqueleto e genes reguladores associados à variação antigênica, sugerindo a conservação dessas regiões em tripanosomas africanos (Sistrom et al., 2014).

#### *1.1.1.2 Cryptosporidium muris*

Os parasitas protozoários do gênero *Cryptosporidium* transmitidos por água ou comida contaminada, completam seu ciclo de vida em um único hospedeiro e contaminam hospedeiros vertebrados (Lumadue et al., 1998). As espécies de *Cryptosporidium* causam gastroenterite aguda e diarreia em todo o mundo. Eles são membros do filo Apicomplexa - protozoários patógenos que invadem células hospedeiras usando um complexo apical especializado e geralmente são transmitidos por um vetor invertebrado ou hospedeiro intermediário. Em contraste com outros Apicomplexa, o *Cryptosporidium* é transmitido pela ingestão de oocistos e completa seu ciclo de vida em um único hospedeiro. Nenhuma terapia está disponível e o controle se concentra na eliminação de oocistos em fontes de água (Xu et al, 2004).

*Cryptosporidium muris*, um patógeno comum em roedores, foi a primeira espécie *Cryptosporidium* descoberta e inicialmente, juntamente com outras espécies que causam criptosporidíase, pensou-se que se limitava apenas aos hospedeiros animais (Tyzzer, 1907). A infecção por *Cryptosporidium* em seres humanos não foi reconhecida até 1976 (Nime et al, 1976). A partir do ano 2000, com o advento de técnicas moleculares, o *Cryptosporidium muris* foi identificado pela primeira vez em humanos, particularmente em indivíduos imunocomprometidos (Guyot et al, 2001), (Tiangtip & Jongwutiwes, 2002), (Palmer et al, 2003).

Esta espécie foi adaptada para viver no estômago, enquanto outras espécies de *Cryptosporidium* se adaptaram para viver nos intestinos. A distância evolutiva entre o *C. muris* e as espécies intestinais *C. parvum* e *C. hominis* o torna um bom candidato a estudos genômicos (NCBI - Genomes, 2018).

No entanto, estudos extensivos de genômica comparativa utilizando espécies do gênero *Cryptosporidium* não foram realizados, exceto pela genotipagem *multilocus* realizada por Xiao e colabores (2000), que não fornece informações suficientes sobre as relações evolutivas entre essas espécies.

#### *1.1.1.3 Entamoeba invadens*

*Entamoeba invadens* é um parasita Amoebozoa, (gênero *Entamoeba)* de répteis e está intimamente relacionado com o parasita humano *Entamoeba histolytica*, causando doença invasiva semelhante (amebíase) em répteis (Ehrenkaufer et al, 2013), além de uma morfologia e ciclo de vida semelhantes

(Sanchez et al, 1994). É geralmente considerado um parasita de quelônios, mas também tem sido implicado como causa de colite, diarreia e morte em gopher (*Gopherus polyphemus*) e tartarugas de leopardo (*Geochelone pardalis*) (Bradford et al., 2008). O genoma do *Entamoeba invadens* tem sido usado como modelo para o estudo do desenvolvimento e encistamento *in vitro*, particularmente devido a dificuldades associadas ao estudo de encistamento em *E. histolytica* (Ehrenkaufer et al, 2013).

## 1.2 Homologia

Proteínas homólogas compartilham uma origem comum, e podem ocorrer na mesma espécie ou em espécies diferentes (Alberts et al., 2014). Entre os diversos eventos relacionados à homologia, podemos destacar: parálogos, ortólogos, genes órfãos, transferência horizontal de genes, perda gênica, entre outros. Dentre esses eventos, os mais comuns são eventos de paralogia, ortologia e seus derivados (Koonin, 2005).

Parálogos e ortólogos são dois tipos de relação homóloga, em que seus genes evoluíram respectivamente, por duplicação e por descendência vertical de um único gene ancestral (Koonin, 2005). Genes ortólogos tendem a manter sua função biológica, diferentemente dos parálogos que podem sofrer mais mutações e divergir na sua função (Alberts et al., 2014). A Figura 1-2 demonstra exemplos das relações de homologia, com ênfase às definições de ortologia e paralogia.

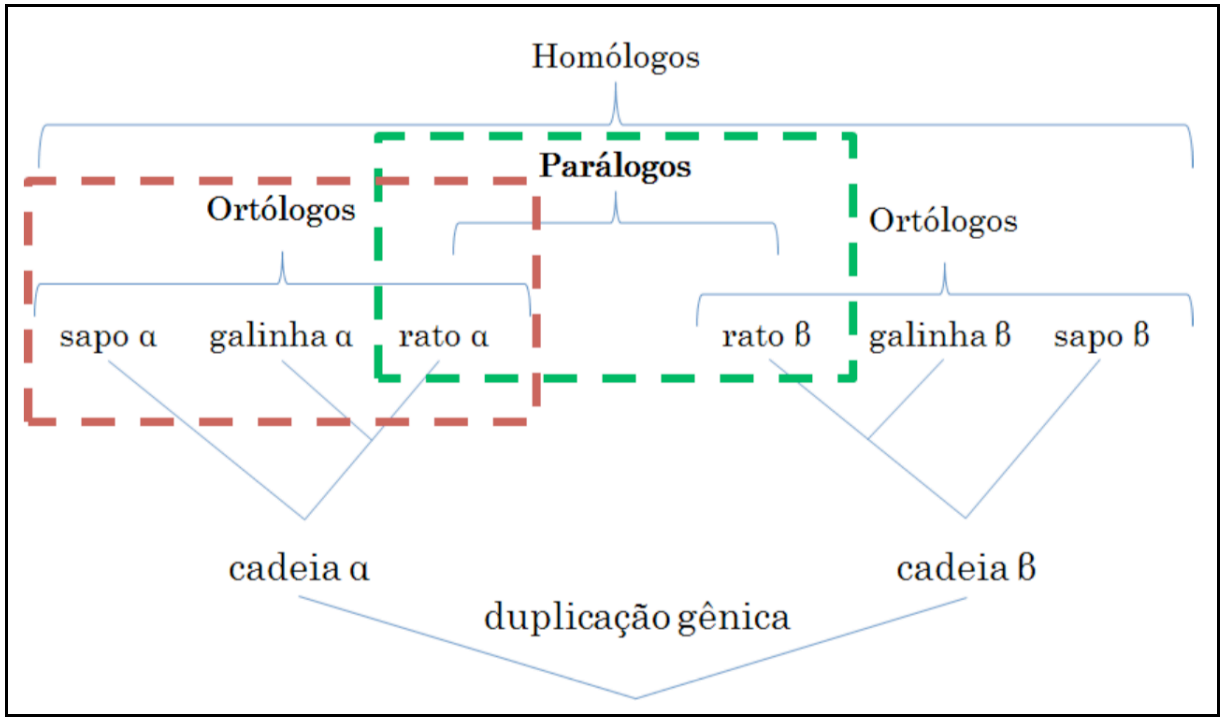



**Figura 1-2: Relações de homologia, paralogia e ortologia utilizando o gene da hemoglobina.**
Ortólogos: Componentes biológicos homólogos (genes, proteínas, estruturas) em diferentes espécies que surgiram de um único componente presente no ancestral comum da espécie; Parálogos: Componentes biológicos homólogos dentro de uma única espécie que surgiram por duplicação gênica. A área retangular vermelha representa um conjunto de genes ortólogos e a área retangular verde, um conjunto de genes parálogos (Fonte: adaptado de https://www.ncbi.nlm.nih.gov/books/NBK62051/).

A genômica comparativa se baseia principalmente no estudo da dinâmica evolutiva dos organismos, seus respectivos genes e proteínas e na homologia, o que se mostra muito útil para aumentar a nossa compreensão sobre a evolução das espécies pela comparação dos seus genomas completos ou de genes específicos de cada espécie (Hardison, 2003).

A crescente disponibilidade de dados de sequências de genomas completos de muitos e diversos organismos tornou possível a observação de características relacionadas à saúde, fenótipos e traços das mesmas, além de analisar a dinâmica evolutiva entre espécies distintas, com foco, por exemplo, na conservação de genes entre elas (Koonin & Galperin, 2003).

A detecção de origem evolutiva comum (homologia) é um meio primário de inferir a estrutura e a função da proteína (Margelevičius & Venclovas, 2010). Para atribuir funções a sequências e estudar sua evolução, estudos comparativos têm

sido utilizados para analisar genomas completos (Kelly et al, 2014; Ehrenkaufer et al, 2013).

### 1.2.1 Métodos para a inferência de homologia

As informações geradas pelos projetos genomas tornam necessário o desenvolvimento de ferramentas computacionais com o objetivo de auxiliar na interpretação dos dados genômicos.

Existem diversos métodos para identificação de homólogos. Algumas técnicas se baseiam na análise de similaridade de sequências de proteínas, outras em alinhamentos múltiplos, outras em Perfil do Modelo Oculto de Markov (pHMM) contra sequências de proteínas, outras em pHMM contra pHMM e outras em filogenia.

#### 1.2.1.1 Métodos para a inferência de homologia: Comparação proteína - proteína

A comparação entre sequências de proteínas é o principal e mais comum meio para inferir a homologia (Margelevičius & Venclovas, 2010).

A análise de similaridade de sequências tem sido usada em diferentes abordagens: motivos, domínios, genes/proteínas inteiras, e até possíveis estruturas secundárias ou terciárias (Eisen & Wu, 2002).

Segundo Koonin & Galperin (2003), a similaridade, principalmente em níveis mais baixos, para ser inferência confiável de homologia, deve atender a um ou mais dos seguintes requisitos:

(i) Estender-se por um trecho considerável da sequência e deve ser estatisticamente significativa, segundo critérios conhecidamente confiáveis;

(ii) Se a similaridade entre as sequências for baixa, o padrão deve ser idêntico e resíduos similares de aminoácidos devem ser encontrados em várias sequências em alinhamentos múltiplos;

(iii) O padrão de similaridade das sequências deve refletir a similaridade entre estruturas experimentalmente determinadas das respectivas proteínas ou ao menos corresponder os elementos chave dessa estrutura.

A homologia pode ser inferida quando 2 sequências compartilham mais similaridade do que seria esperado ao acaso. Quando a alta similaridade é observada, a explicação mais simples é que essas 2 sequências não surgiram independentemente, elas descendem de um ancestral em comum (Pearson, 2013). Validação da homologia inferida por meio de similaridade entre as sequências é sempre desejável, seja pela inferência de domínios conservados e/ou motivos.

Algumas ferramentas populares, como os pacotes FASTA (Pearson, 1990) e BLAST (Altschul et al., 1990), são implementações dessa abordagem.

Outros programas utilizam scores de alinhamentos com melhores hits recíprocos, como o InParanoid (Remm at al., 2001), o OrthoSearch (da Cruz et al., 2008) e o OrthoMCL (Li et al., 2003). Esses scores são obtidos utilizando uma matriz de pontuações de similaridade para todos os possíveis pares de resíduos, sendo que identidades e substituições conservadas têm pontuações positivas, enquanto substituições improváveis têm pontuações negativas (Altschul et al., 1990).

Existem ainda programas que trabalham com cálculo de distância evolutiva como parte do seu método para inferência de homologia, como o RSD (Wall & DeLuca, 2007) e o OMA (Dessimoz et al., 2005). Estimativas de distâncias evolutivas também podem ser inferidas via comparação entre sequências alinhadas utilizando matrizes de pontuações de similaridade (Sonnhammer & Hollich, 2005). A distância evolutiva entre duas sequências alinhadas comumente utiliza como padrão de medida o PAM ou “Percent Accepted (point) Mutation”, termo introduzido por Dayhoff et al (1978), onde o autor propõe o uso de matrizes de pontuação para obtenção de escores de alinhamentos.

#### 1.2.1.2 Métodos para a inferência de homologia: Comparação pHMM - proteína

Para proteínas estreitamente relacionadas, a similaridade da sequência pode ser detectada facilmente, no entanto, a semelhança torna-se fraca e difícil de distinguir à medida que a distância evolutiva aumenta (Margelevičius & Venclovas, 2010). Sequências homólogas nem sempre compartilham similaridade significativa, é comum a existência de alinhamentos de proteínas homólogas que não são significativos, mas são claramente homólogos com base em semelhança estrutural estatisticamente significativa ou forte similaridade de sequência com uma sequência intermediária (Pearson, 2013).

O uso de perfis do modelo oculto de Markov (pHMM) representa um importante avanço em termos de sensibilidade de busca de sequencias para detecção de homologias distantes (Wheeler & Eddy, 2013). O pHMM é uma representação condensada dos alinhamentos múltiplos e contém para cada resíduo a probabilidade de observar eventos evolutivos (inserção, mutação e exclusão) (Remmert et al., 2011).

Existem programas com uma maior sensibilidade que o BLAST e FASTA, tais como o HMMER (Wheeler & Eddy, 2013), o SAM (Hughey & Krogh, 1996), o THMM (Qian & Goldstein, 2004), e o HH-suite (Remmert et al., 2011), todos baseados no algoritmo HMM (Hidden Markov Model) e utilizam esse tipo especial de modelo probabilístico, construído a partir de um conjunto de sequências homólogas e avaliam o quanto outras sequências se assemelham ao perfil.

Métodos que utilizam pHMM têm demonstrado uma alta eficiência na detecção de homologias distantes entre sequências (Gough et al., 2001; Madera et al., 2002; Park et al. 1998; Wistrand & Sonnhammer, 2005).

Além do pHMM, existe ainda outro tipo de perfil também utilizado para a comparação “perfil - proteína”, chamado Matriz de Posição (position-specific score matrices – PSSM ou Position Weight Matrices - PWM) usado em ferramentas como PSI-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) que também tem demonstrado uma boa eficiência na detecção de homologias distantes comparados com os métodos que utilizam a comparação “proteína x proteína” (Altschul et al., 1997). Um PWM especifica a distribuição de frequência de nucleotídeos em cada posição dos sítios de ligação (Stormo & Fields, 1998) e é considerado mais simples e com menos parâmetros que o pHMM, porém o PWM não pode modelar motivos de comprimento variável e não permite a modelagem de inserções e deleções de nucleotídeos dentro de um sítio de ligação, sendo isto uma vantagem do pHMM pois ele permite a modelagem de um modelo de motivos mais abrangente do que os PWMs (Sinha, 2006).

#### 1.2.1.3 Métodos para a inferência de homologia: Comparação pHMM - pHMM

A detecção de homologia distante é uma abordagem computacional amplamente utilizada para estudar a evolução, a estrutura e a função das proteínas (Margelevičius, Laganeckas & Venclovas, 2010).

Segundo Remmert e colaboradores (2011), quando o objetivo é procurar homólogos distantes, a melhor estratégia é usar a maior quantidade de informação quanto possível nos dois lados que serão comparados, para assim poder distinguir melhor entre resultados positivos e falsos-positivos na produção de alinhamentos. Ferramentas que fazem comparação pHMM - pHMM representam os métodos mais sensíveis existentes hoje (Remmert et al., 2011).

Pacotes de programas como o COMA (Margelevičius & Venclovas, 2010) e HHsearch, presente no pacote HH-suite (Remmert et al., 2011) propiciam a comparação entre pHMMs. Esse tipo de comparação tem demonstrado ainda mais eficiência na detecção de homologias distantes que a comparação "pHMM - proteína" (Remmert et al., 2011).

Para Wilson e colaboradores (2009), a identificação de homólogos distantes resultaria na identificação de superfamílias de proteínas (Figura 1-3).

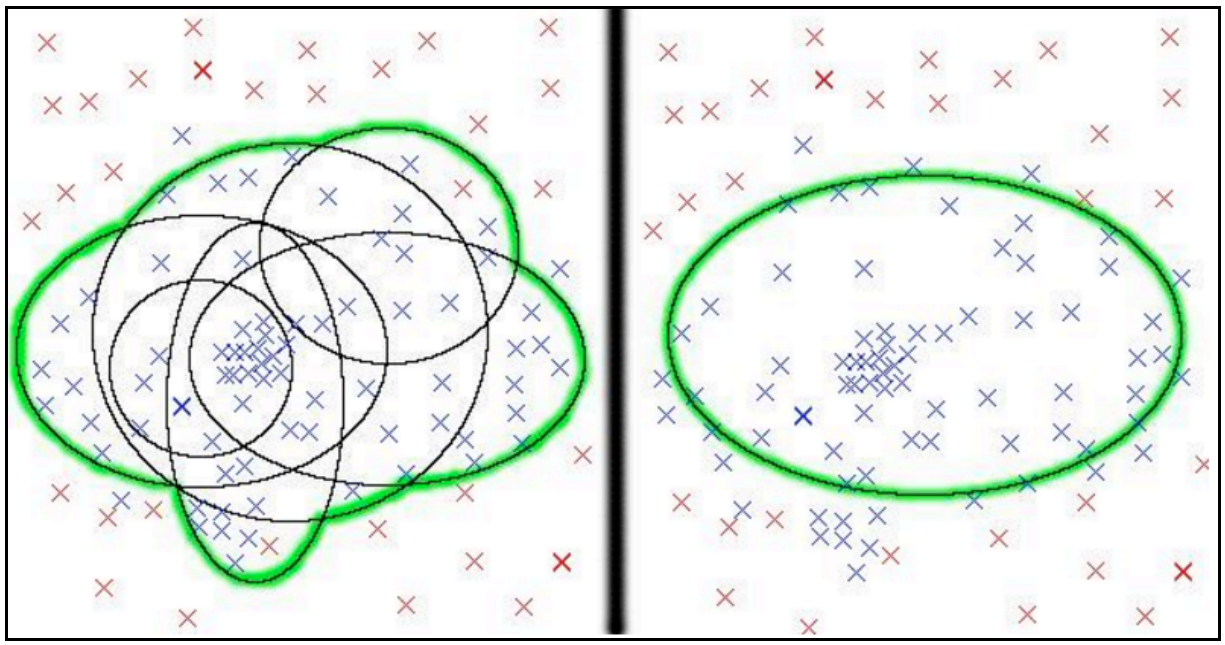


**Figura 1-3: Representação de superfamílias usando múltiplos HMM.**
Este diagrama mostra 2 exemplos de modelos em um espaço de sequência imaginário. Cada "X" representa uma sequência de estrutura desconhecida, as cruzes azuis são membros da superfamília e as cruzes vermelhas não são. Cada elipse representa um único modelo, o esboço verde representa a superfamília como a união de vários modelos, que é a maneira pela qual eles são tratados pelo SUPERFAMILY.

(Fonte: http://supfam.cs.bris.ac.uk/SUPERFAMILY/representation.html)

#### 1.2.1.4 Filogenia

A filogenia se baseia no princípio de que, se os genomas evoluem por uma acumulação gradual de mutações, então as diferenças entre as sequências de nucleotídeos de 2 genomas devem conseguir indicar o quão recentemente esses genomas compartilharam um ancestral comum. Se as diferenças são muitas, o ancestral comum deve ser mais antigo do que em genomas com poucas diferenças. Então, quando comparamos 3 ou mais genomas, deve ser possível identificar as relações evolucionárias entre eles (Brown, 2002).

A filogenia analisa a informação genética no contexto de sua evolução ao invés de utilizar apenas similaridade de sequências e é considerada uma abordagem promissora para inferir relações de homologia (Silva et al., 2012; Gabaldón, 2008; Eisen & Wu, 2002).

Uma árvore filogenética descreve as relações evolutivas entre as sequências e espécies envolvidas, de modo que os eventos de especiação e duplicação possam então ser mapeados nos nós desta árvore (Gabaldón, 2008).

Estudos utilizando essas técnicas foram utilizadas em genomas completos de protozoários, como por exemplo *Entamoeba histolytica* (Cuadrat et al,. 2014), *Trypanosoma cruzi* (Jackson et al., 2016) e *Cryptosporidium hominis* (Xu, Widmer & Wang, 2004).

### 1.2.2 Bases de dados de homólogos

Com a finalidade de facilitar a anotação funcional de novos genomas, diversas bases de dados de homólogos foram criadas, como por exemplo KEGG, eggNOG e OrthoMCL-DB. Cada uma destas bases de dados se baseia em algoritmos e/ou pipelines distintos de inferência de homologia.

#### KEGG Orthology (KO)

O Kyoto Encyclopedia of Genes and Genomes - KEGG (Ogata et al., 1998) é uma base de dados para a compreensão de funções de alto nível de sistemas biológicos como a célula, o organismo e o ecossistema, a partir de informações de nível molecular, especialmente conjuntos de dados moleculares de grande escala gerados pelo sequenciamento do genomas e outros experimentos tecnológicos de

alto rendimento. A base de dados KO é parte do sistema KEGG e seus grupos ortólogos foram criados de forma manual, por análise de similaridade de sequências e análise funcional. Seus membros devem ter complexo molecular comum ou pertencer a uma via metabólica conservada. O agrupamento de seus ortólogos é feito por um algoritmo que compara grafos de forma heurística, genoma contra genoma e genoma contra via metabólica.

**eggNOG**

O evolutionary genealogy of genes: Non-supervised Orthologous Groups - eggNOG (Powell et al., 2014) é uma base de dados de grupos ortólogos e de anotação funcional hospedado pelo EMBL (European Molecular Biology Laboratory). Baseia-se na ideia original de COGs (Clusters of Orthologous Groups) (Tatusov et al., 2000) e expande essa ideia a grupos ortólogos não supervisionados construídos a partir de um grande número de organismos (Powell et al., 2012).

As anotações funcionais para cada proteína são realizadas de forma automática, através de heurística que busca descrições em texto livre de outras bases de dados (KOG, COG, KO, etc), no caso desta busca não retornar nenhum dado, é realizada buscas através do Gene Ontology (Ashburner et al., 2000).

Seus grupos ortólogos também possuem parálogos, inferidos através de modelos de Markov criados a partir de marcadores genéticos universais curados (Ciccarelli et al., 2006).

**OrthoMCL-DB**

O OrthoMCL-DB (Chen et al., 2006) é uma base da dados pública com grupos homólogos inferidos pelo programa OrthoMCL e contém grupos ortólogos para eucariotos e procariotos sequenciados. O OrthoMCL-DB fornece uma grande variedade de funcionalidades, incluindo arquitetura de domínio para cada grupo, padrões filogenéticos para cada grupo e consultas avançadas, incluindo pesquisas de padrões filogenéticos. Os grupos homólogos do OrthoMCL-DB possuem proteínas ortólogas e parálogas, inferidas com abordagem de melhores hits recíprocos do BLAST e por cadeias de Markov (Enright, Dongen & Ouzounis, 2002).

### 1.2.3 Outras bases de dados biológicos

Além das bases de dados de homólogos citadas acima, outras bases de dados também foram criadas visando facilitar a transferência de anotação funcional, como o SUPERFAMILY, o CDD e o Pfam.

#### SUPERFAMILY

O SUPERFAMILY é uma base de dados de anotação estrutural e funcional para proteínas e genomas. A anotação no SUPERFAMILY é baseada em uma coleção de modelos ocultos de Markov (HMMs), que representam domínios estruturais de proteínas no nível de superfamília disponíveis no banco de dados da iniciativa Classificação Estrutural das Proteínas ou SCOP (Wilson et al., 2009).

O SCOP fornece uma descrição detalhada e abrangente das relações das estruturas protéicas conhecidas. A classificação está em níveis hierárquicos: os dois primeiros níveis, família e superfamília, descrevem relações evolutivas próximas e distantes (Hubbard et al., 1997).

#### CDD

O Conserved Domain Database - CDD (Marchler-Bauer et al., 2015) é um recurso de anotação de proteínas que consiste em uma coleção de modelos de alinhamentos múltiplos de sequências bem anotadas para domínios antigos e proteínas completas. Estes modelos estão disponíveis como matrizes de pontuação específicas de posição (PSSMs) para identificação rápida de domínios conservados em sequências de proteínas via RPS-BLAST. A base de dados do CDD inclui domínios com curadoria NCBI, que usam informações de estrutura 3D para definir explicitamente limites de domínio e fornecer informações sobre relacionamentos de sequência, estrutura e função, bem como modelos de domínio importados de uma série de bases de dados externos (Pfam, SMART, COG, PRK, TIGRFAM).

#### Pfam

A base de dados Pfam (Finn et al., 2016) é uma coleção de famílias de proteínas, cada uma representada por alinhamentos de sequências múltiplas e modelos ocultos de Markov (HMMs). O Pfam também gera agrupamentos de nível superior de entradas relacionadas, conhecidas como clans. Um clan Pfam é uma

coleção de entradas Pfam (famílias de proteínas representadas por domínios conservados) que são relacionadas por similaridade de sequência, estrutura ou pHMM.

## 1.3 Organização da tese

Esta tese possui como tema central o estudo da genômica comparativa em ambiente computacional de espécies evolutivamente distantes de protozoários. Com 3 contribuições relevantes para a genômica comparativa:

(i) Comparação entre dois programas de inferência de homologia: OrthoMCL (baseado em scores de alinhamentos com melhores hits recíprocos do Blast) e OMA (baseado em Smith-Waterman paralelizado e distância evolutiva).

(ii) Uma novo método para inferência de homólogos, baseada na reconciliação de grupos homólogos distintos.

(iii) A identificação de homólogos distantes em protozoários de diferentes filos.

Todos os resultados obtidos nesta tese são explorados nos capítulos a seguir.

- O capítulo 2 apresenta o objetivo geral e os objetivos específicos desta tese;

O capítulo 3 apresenta os materiais e métodos utilizados no desenvolvimento dos experimentos realizados. Detalhadamente:

- O Capítulo 3.1 descreve o experimento no qual foram inferidos homólogos para *Cryptosporidium muris*, *Entamoeba invadens* e *Trypanosoma grayi*, utilizando os programas OrthoMCL e OMA, além da comparação entre seus resultados.
- O Capítulo 3.2 descreve a metodologia aqui proposta de inferência de grupos homólogos com base na reconciliação dos resultados das 2 metodologias distintas do Capítulo 3.1.
- O Capítulo 3.3 descreve o experimento que teve como objetivo identificar homólogos distantes por meio de comparação pHMM – pHMM usando como base os resultados apresentados no Capítulo 3.1.

- O capítulo 4 e seus subtópicos apresentam os resultados obtidos nestas atividades.
- O Capítulo 5 é composto pela discussão dos resultados alcançados.
- O Capítulo 6 apresenta as conclusões sobre esta tese.

# 2 OBJETIVOS

## 2.1 Objetivo geral

Realizar estudos de homologia visando a identificação de homólogos distantes através de abordagens de reconciliação de programas de inferência de homologia, em comparação com a abordagem pHMM – pHMM.

## 2.2 Objetivos específicos

2.2.1 Identificar homólogos distantes entre 3 espécies de protozoários distantes utilizando dois programas para inferência de homologia com diferentes metodologias, além de realizar comparações entre seus resultados.

2.2.2 Desenvolver um método para inferência de homologia baseada na reconciliação de grupos homólogos inferidos por duas abordagens distintas.

2.2.3 Identificar homólogos distantes através da comparação pHMM - pHMM e validá-los com a caracterização de superfamílias na base de dados SUPERFAMILY.

# 3 MATERIAL E MÉTODOS

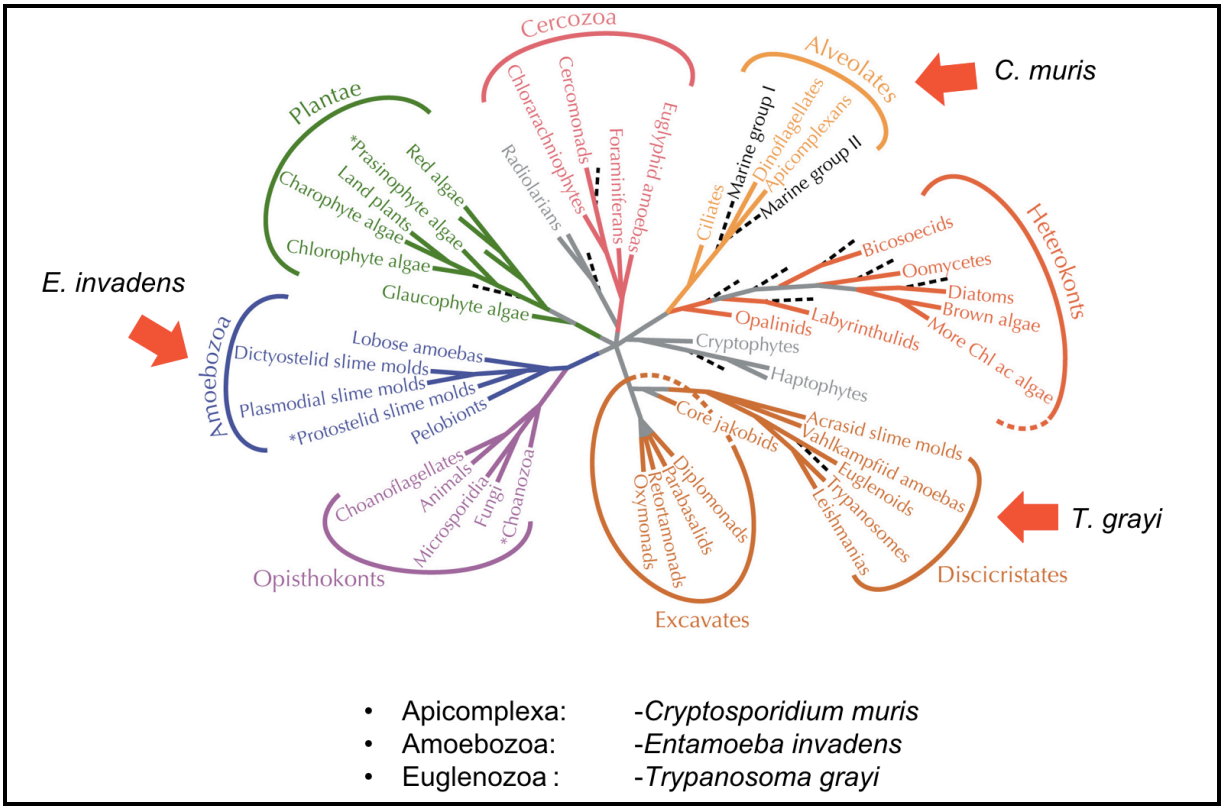
## 3.1 Inferência de grupos homólogos em protozoários evolutivamente distantes com os programas OrthoMCL e OMA.

### 3.1.1 Caracterização e processo de criação de solução do estudo

Esta contribuição é diretamente associada ao objetivo 2.2.1 desta tese.

Este experimento teve como objetivo inferir genes homólogos entre protozoários evolutivamente distantes, utilizando diferentes metodologias e fazer uma análise e comparação entre seus resultados.

Uma vez que, até onde sabemos, as 3 seguintes espécies de Protozoários não foram usadas para a genômica comparativa antes, e queríamos realizar testes sobre a inferência de grupos homólogos em espécies de protozoários distantes, nós fizemos estudos utilizando espécies pertencentes a diferentes filos: *Cryptosporidium muris* (Apicomplexa), *Entamoeba invadens* (Amoebozoa) e *Trypanosoma grayi* (Euglenozoa) (Figura 3-1). A Figura 3-2 mostra um fluxograma detalhando a metodologia desse experimento.




- Apicomplexa: -*Cryptosporidium muris*
- Amoebozoa: -*Entamoeba invadens*
- Euglenozoa : -*Trypanosoma grayi*

**Figura 3-1: Árvore filogenética dos maiores grupos evolucionários de eucariotos.**
Filogeneticamente, os 3 protozoários utilizados nessa tese estão espalhados na árvore da vida, estando presentes nas linhagens discicristados, alveolatas e amebozoas. (Fonte: adaptado de Baldauf, 2003).

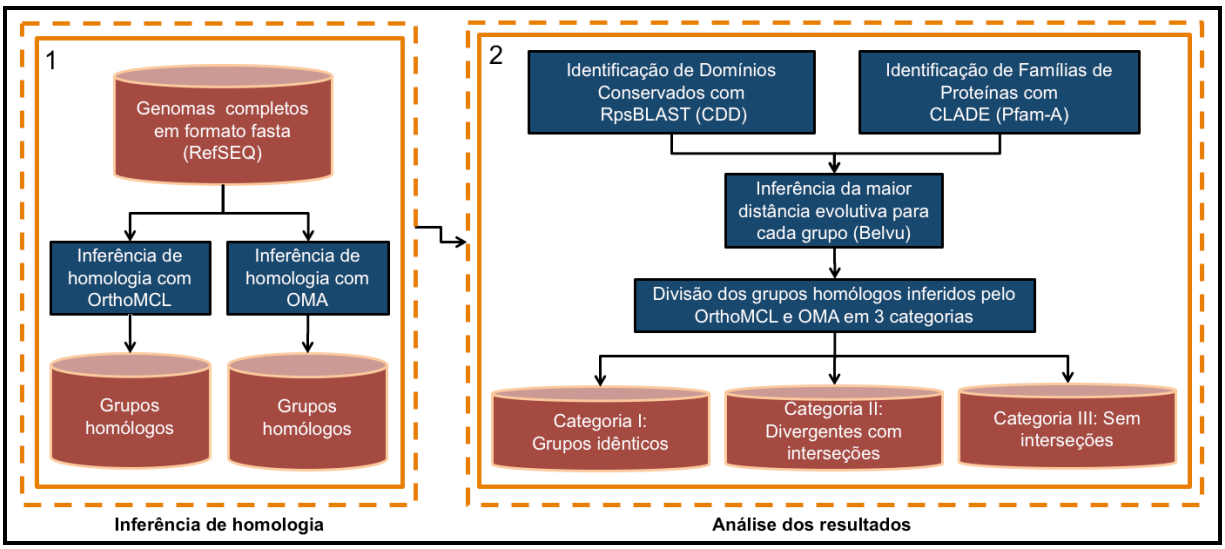



**Figura 3-2: Fluxograma que detalha o processo de inferência de grupos homólogos do OrthoMCL e do OMA e realização de análises.**

### 3.1.2 Genomas completos utilizados

Os genomas completos utilizados nesta tese foram obtidos no RefSeq/NCBI (ftp.ncbi.nlm.nih.gov/genomes/refseq/protozoa) em formato fasta, em detalhes:

O conjunto completo dos dados utilizados tem 26.514 proteínas, destas 14,83% (3.934/26.514) são de *C. muris*, 45,24% (11.997/26.514) de *E. invadens* e 39,91% (10.583/26.514) de *T. grayi* (Tabela 1).

**Tabela 1: Genomas completos utilizados nos experimentos da tese.**
O genoma completo de *Cryptosporidium muris*, *Entamoeba invadens* e *Trypanosoma grayi*, obtidos da RefSeq / NCBI, com suas respectivas versões.

| Organismo | Versão do Genoma | Nº de proteínas |
|---|---|---|
| *Cryptosporidium muris* | GCF_000006515.1_JCVI_cmg_v1.0 | 3.934 |
| *Entamoeba invadens* | GCF_000330505.1_EIA2_v2 | 11.997 |
| *Trypanosoma grayi* | GCF_000691245.1_Tgr_V1 | 10.583 |

### 3.1.3 Identificação de homólogos através dos programas OMA e OrthoMCL

Foram utilizados 2 programas para identificar genes homólogos: OMA (infere grupos ortólogos) e OrthoMCL (infere grupos ortólogos e parálogos).

O OrthoMCL foi obtido no site orthomcl.org (http://orthomcl.org/common/downloads/software/) e usado em conjunto com o BLAST versão 2.5.0+ (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). Utilizamos e-value de 1E-05 como valor de corte, de acordo com o protocolo descrito por Coutinho et al, 2011.

O OMA versão 1.0.1 foi obtido em omabrowser.org (http://omabrowser.org/standalone/) e foi executado com os parâmetros padrão.

### 3.1.4 Análise dos resultados do OMA e OrthoMCL

Os grupos homólogos inferidos por OrthoMCL e OMA foram divididos em 3 categorias, de acordo com o grau de concordância entre eles, a saber: "Grupos idênticos" (denominados a partir de agora como Categoria I), formados por grupos homólogos onde os dois programas concordaram, contendo exatamente as mesmas proteínas; "Divergentes com as interseções" (denominados a partir de agora como Categoria II), formado por grupos homólogos que não são idênticos entre os dois programas, mas compartilham pelo menos uma proteína em comum; e "Sem intersecções" (denominados a partir de agora como Categoria III), formada por grupos homólogos onde nenhuma proteína é compartilhada entre os resultados dos dois programas.

### 3.1.5 Inferência de distância evolutiva do OMA e OrthoMCL

Para calcular a distância evolutiva para cada grupo homólogo, utilizamos o programa Belvu (Sonnhammer & Hollich, 2005) versão "Ubuntu 12.04.3 64bit" (http://sonnhammer.sbc.su.se/download/software/belvu/), com os seguintes parâmetros: “Tree options: Use Scoredist distance correction (default)” and “Print distance matrix and exit”.

Para verificar se as distribuições das amostras são estatisticamente diferentes foi utilizado o programa R e realizado o teste Wilcoxon-Mann-Whitney, uma vez que as amostras não apresentaram uma distribuição normal.

Com o intuito de comparar a distância máxima encontrada em cada grupo homólogo, separamos os grupos OrthoMCL sem parálogos, já que o OMA não inferiu parálogos.

### 3.1.6 Identificação de domínios conservados (CDD)

RPS-BLAST versão 2.2.13 (https://www.ncbi.nlm.nih.gov/) foi utilizado para detectar domínios conservados contra a base de dados de domínios conservados CDD (Conserved Domain Database) – versão CDD.v3.12 (https://www.Ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). Também foi utilizado e-value de 1E-05 como valor de corte e o programa foi executado com seus parâmetros padrão.

### 3.1.7 Identificação de famílias de proteínas (Pfam-A)

Para obter as predições de famílias de proteínas da base de dados Pfam-A, utilizamos a ferramenta CLADE (Bernardes et al., 2015; 2016). O programa e sua biblioteca modelo podem ser baixadas no seu portal (http://www.lcqb.upmc.fr/CLADE). Também usamos e-value de 1E-05 como valor de corte e o programa foi executado com seus parâmetros padrão.

## 3.2 Inferência de homologia baseada em uma abordagem de reconciliação para a genômica comparativa de protozoários.

### 3.2.1 Caracterização e processo de criação de solução do estudo

Esta contribuição é diretamente associada ao objetivo 2.2.2 desta tese.

Este experimento usou como base de dados os homólogos inferidos pelos programas OrthoMCL e OMA pertencentes a Categoria II: "Divergentes com as interseções" no capítulo 3.1 desta tese e teve como objetivo identificar grupos homólogos que não pudessem ser inferidos separadamente por qualquer um dos 2 programas de inferência de homologia utilizados (OMA e OrthoMCL), utilizando uma abordagem de reconciliação dessas diferentes metodologias. Estes novos grupos homólogos inferidos com nossa abordagem de reconciliação foram chamados de **Supergrupos** homólogos.

Neste trabalho, denominamos Supergrupos homólogos aos grupos formados pela concatenação de todos os grupos homólogos com proteínas em comum (Categoria II) que tiveram um aumento no seu número de proteínas em relação a todos os grupos homólogos (OrthoMCL e OMA) que os originaram e que, além disso, tiveram identificado mesmo domínio conservado (CDD) ou a mesma família de proteínas (Pfam-A) em todas as suas proteínas.

A Figura 3-3 mostra um fluxograma detalhando a metodologia desse experimento.

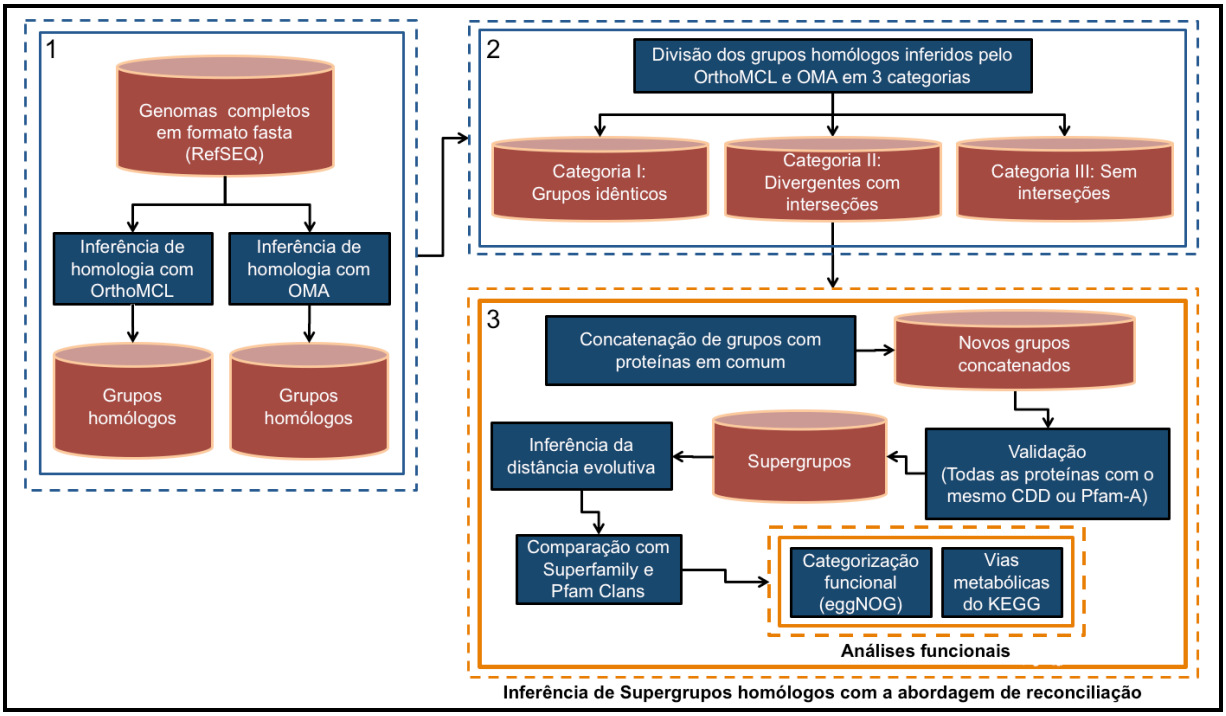



**Figura 3-3: Fluxograma que detalha o processo de inferência dos Supergrupos e suas análises funcionais.**

### 3.2.2 Algoritmo de reconciliação – Inferência de Supergrupos

O algoritmo de reconciliação (arquivo adicional 1) concatenou grupos homólogos do OrthoMCL e do OMA com proteína(s) em comum, de forma recursiva, até que não fossem mais encontrados grupos com proteína(s) em comum entre si (Figura 3-4).

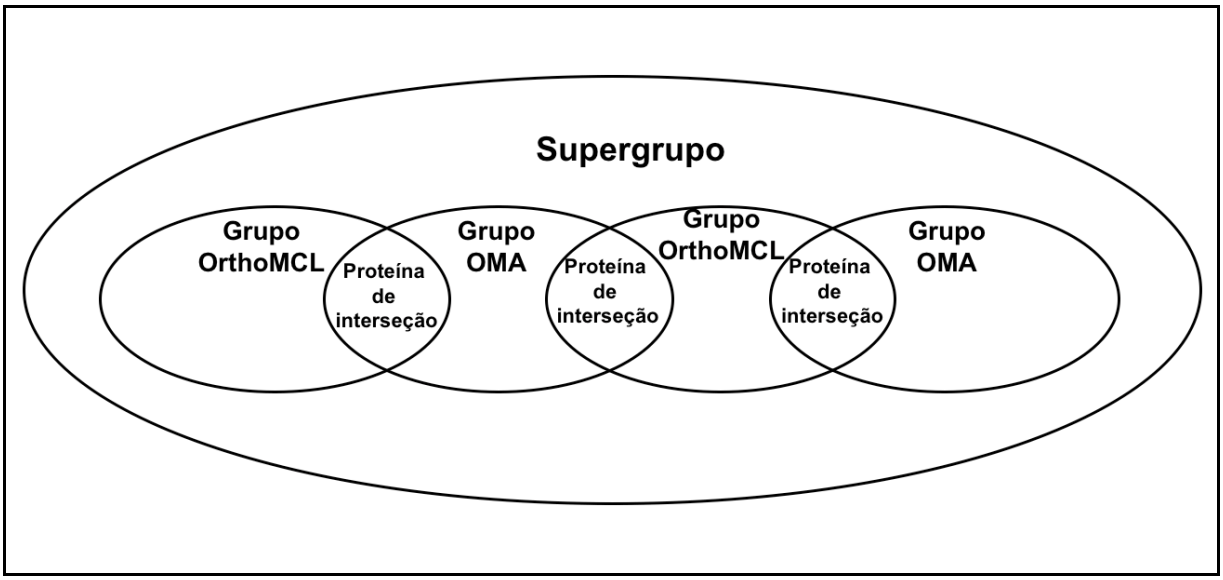


**Figura 3-4: O algoritmo de reconciliação.**
Método para inferir Supergrupos homólogos usando a reconciliação dos resultados OrthoMCL e OMA pertencentes a categoria II: "Divergentes com interseções", onde os grupos homólogos possuíam pelo menos uma proteína em comum.

### 3.2.3 Validação por CDD ou Pfam-A

A validação dos nossos resultados foi feita verificando se todas as proteínas dos grupos concatenados possuem o mesmo domínio conservado (CDD) ou pertencem a mesma família de proteínas (Pfam- A).

### 3.2.4 Filogenia

RAxML, versão 8 (https://sco.h-its.org/exelixis/software.html) foi utilizado para inferir o melhor modelo evolutivo, bem como as árvores com base na máxima verossimilhança, utilizando 500 repetições de bootstrap. O programa foi executado com parâmetros padrão.

### 3.2.5 Comparação com bases de dados SUPERFAMILY e Pfam Clans

Visando confirmar a validação por CDD ou Pfam-A, as proteínas de cada Supergrupo inferido foram mapeadas nas bases de dados SUPERFAMILY e Pfam Clans usando como corte um e-value 1E-05.

### 3.2.6 Análise da conservação das sequências

Para avaliar a conservação de sequências, foram gerados alinhamentos múltiplos para cada um dos grupos inferidos por concatenação da Categoria II deste estudo, usando o programa Mafft, versão 7.271, com seus parâmetros padrão. O programa Alistat, disponível no Hmmer versão 3.0, foi usado para gerar as estatísticas dos alinhamentos múltiplos.

### 3.2.7 Inferência da distância evolutiva nos Supergrupos

Para calcular a maior distância evolutiva para cada um dos Supergrupos homólogos inferidos neste estudo foi utilizado a mesma metodologia apresentada no item 3.1.5 desta tese.

### 3.2.8 Categorização funcional

A categorização funcional foi realizada para os Supergrupos inferidos neste estudo usando análise de similaridade com Hmmer (Wheeler & Eddy, 2013) versão 3.0 (http://eddylab.org/software/hmmer3/3.0/) contra a Base de Dados de Genes Ortólogos - eggNOG (Powell et al., 2014) versão 4.0 (ftp://eggnog.embl.de/eggNOG/4.0/). Para inferir a qual categoria funcional cada proteína pertence, um e-value de 1E-05 como valor de corte. O programa foi executado com seus parâmetros padrão.

### 3.2.9 Vias metabólicas do KEGG

As vias metabólicas foram identificadas para os Supergrupos inferidos neste estudo, usando análise de similaridade com BLASTP (Altschul et al., 1990) (proteína-proteína BLAST) versão 2.5.0+ (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.5.0/) contra a base de dados de eucariotas e genes procariotas da Enciclopédia de Genes e Genomas de Quioto ou “Database of Eukaryotes and Prokaryotes Genes of the Kyoto Encyclopedia of Genes and Genomes” - KEGG versão "setembro de 2016" (ftp://ftp.bioinformatics.jp/kegg/genes/fasta/). Também usamos e-value de 1E-05 como corte e o programa foi executado com parâmetros padrão.

## 3.3 Inferência e análise de homólogos distantes em Protozoa por meio da comparação pHMM - pHMM (perfis de Modelo Oculto de Markov) visando a identificação de superfamílias

### 3.3.1 Caracterização e processo de criação de solução do estudo

Esta contribuição é diretamente associada ao objetivo 2.2.3 desta tese.

Este estudo teve como objetivo inferir homólogos distantes e validá-los usando a base de dados SUPERFAMILY (supfam.org) para identificar superfamílias de proteínas entre os grupos homólogos distantes com melhores hits recíprocos, inferidos por meio da comparação pHMM – pHMM (COMA).

### 3.3.2 Comparação pHMM - pHMM com o COMA

O pacote de programas COMA versão 1.10 (http://www.bti.vu.lt/en/departments/department-of-bioinformatics/software/coma) foi utilizado nesta etapa. Primeiramente, utilizando o programa MakePro (presente no pacote COMA), foi criado um pHMM para cada um dos grupos homólogos inferidos pelo OMA (1.231 grupos) e pelo OrthoMCL (3.092 grupos) (detalhes no capítulo 3.1) e em seguida foram feitas as comparações pHMM - pHMM entre eles utilizando o programa COMA, visando a identificação de homólogos distantes entre os grupos homólogos inferidos previamente (Figura 3-5).

Como estudo de caso, foram utilizados os grupos homólogos que tiveram melhor hit recíproco, conforme apontado por um script feito na linguagem Ruby, escrito para identificar os grupos alvo.

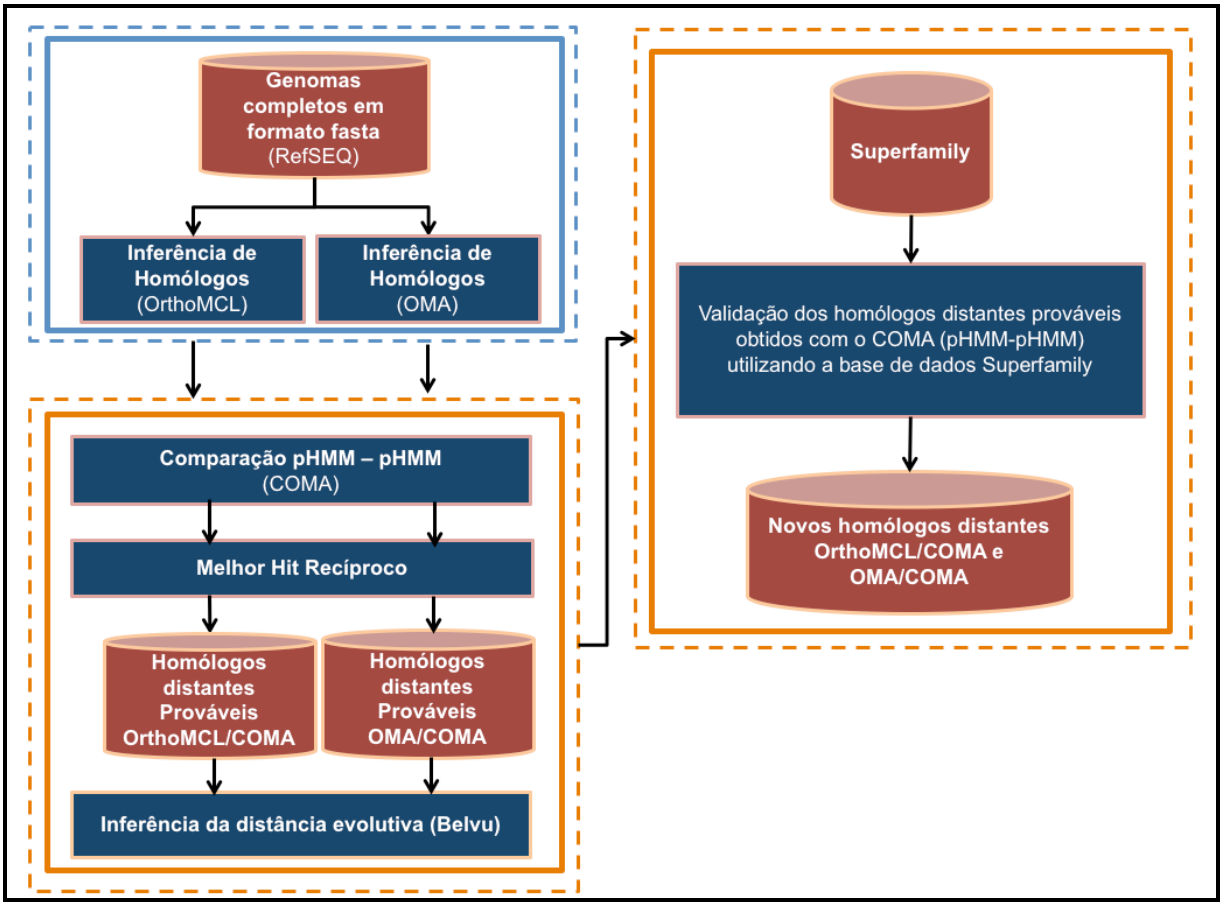



Figura 3-5: Fluxograma que detalha o processo de inferência de homólogos distantes e sua validação via SUPERFAMILY.

### 3.3.3 Validação dos homólogos distantes inferidos pelo COMA

Com o objetivo de validar os grupos homólogos distantes que tiveram melhores hits recíprocos, foram obtidas predições de superfamílias de proteínas feitas na ferramenta própria da base de dados SUPERFAMILY, disponível em seu portal online (http://supfam.org/SUPERFAMILY/), foi utilizado um scritp feito em linguagem Ruby para se obter os resultados destas inferências (arquivo adicional 2).

### 3.3.4 Inferência da distância evolutiva para os homólogos distantes inferidos via pHMM - pHMM

Para calcular a maior distância evolutiva para cada um dos novos homólogos inferidos neste estudo, foi utilizada a mesma metodologia apresentada no item 3.1.5 desta tese.

### 4.3.3 Criação de grafos para análises adicionais

Foram criados grafos representando as relações entre os 3 melhores hits (quando possível) dos grupos homólogos com 3 organismos que apresentaram melhor hit recíproco pHMM – pHMM, utilizando a ferramenta Cytoscape versão 3.6.1 (disponível em http://www.cytoscape.org).

# 4 RESULTADOS

## 4.1 Inferência de grupos homólogos em protozoários evolutivamente distantes com os programas OrthoMCL e OMA

### 4.1.1 Inferência de homologia

A porcentagem de proteínas de *C. muris*, *E. invadens* e *T. grayi* contribuindo para os grupos homólogos inferidos pelo OrthoMCL e pelo OMA é mostrada na Figura 4-1.

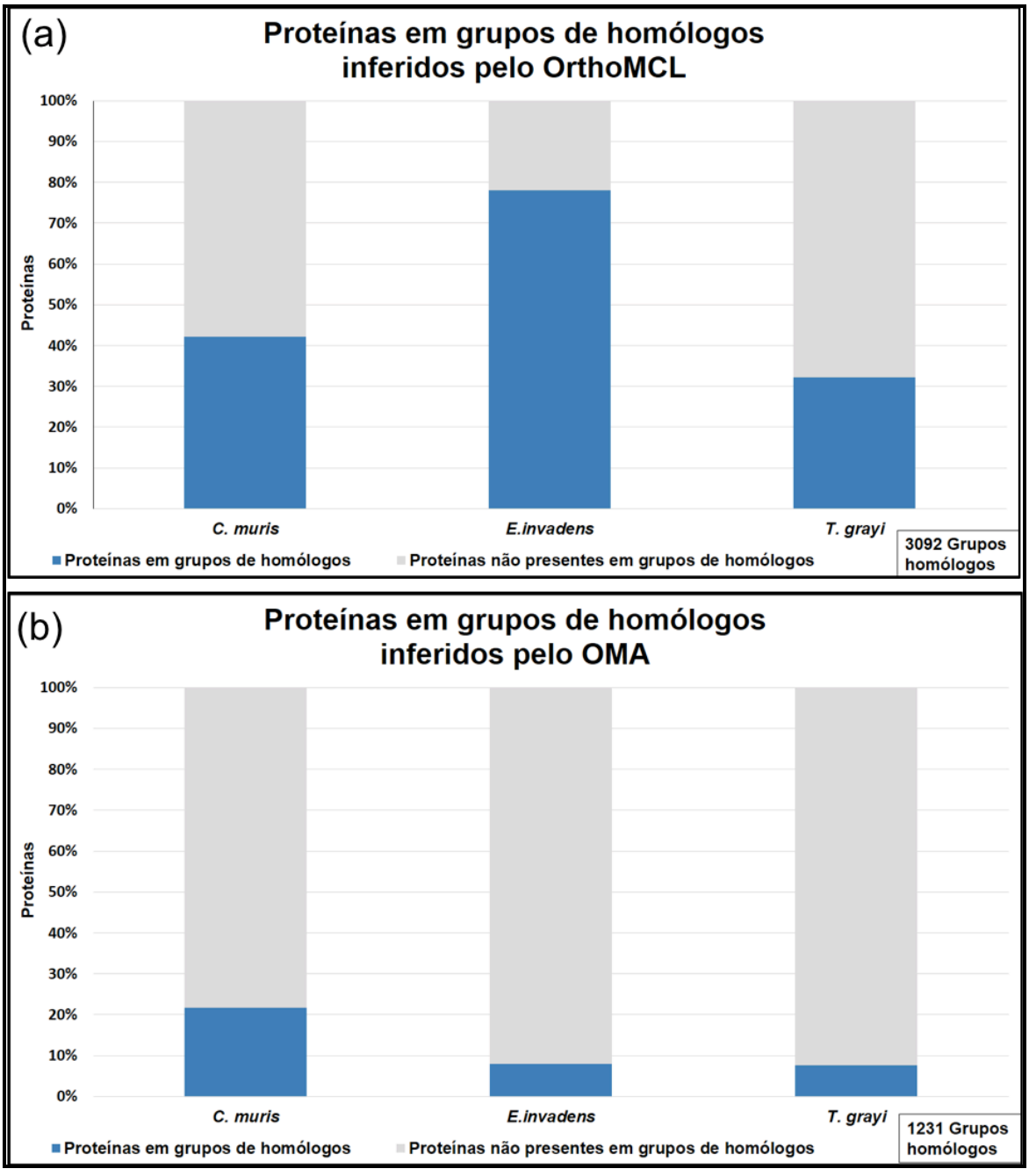


**Figura 4-1: Proteínas de cada espécie contribuindo para os grupos homólogos.** (a) Nos resultados do OrthoMCL, *T. grayi* apresentou menor porcentagem de homólogos com 36,23% (3.835/10.583), seguido de *C. muris* com 42,22% (1.661/3.934) e *E. invadens*, com 78,08% (9.368/11.997) (b) Nos resultados do OMA, *E. invadens* mostrou a menor porcentagem de homólogos com 8,03% (964/11.997), seguido de *T. grayi* com 9,71% (1.028/10.583) e *C. muris* com 21,65% (852/3.934).

O OrthoMCL inferiu 3.092 grupos homólogos, cobrindo 56,06% (14.864/26.514) das proteínas do dataset. Desses grupos, 52,1% (1.611/3.092) são grupos parálogos, sendo 32,14% (994/3.092) parálogos de *E. invadens*, 16,26% (503/3.092) de *T. grayi* e apenas 3,68% (114/3.092) são formados por parálogos de *C. muris*. Além disso, o resultado do OrthoMCL mostrou que 47,89% (1.481/3.092) são grupos ortólogos, sendo 15,97% (494/3.092) grupos sem parálogos (exclusivamente ortólogos) e 31,92% (987/3.092) são grupos homólogos com proteínas ortólogas e parálogas. Entre os grupos homólogos inferidos pelo OrthoMCL, observou-se que: 24,51% (758/3.092) são compartilhados pelas 3 espécies, enquanto que 23,38% (723/3.092) dos grupos são compartilhados por 2 espécies, mostrando a seguinte distribuição: 5.3% (164/3.092) são grupos compartilhados por *E. invadens* e *C. muris*, 7,92% (245/3.092) são grupos compartilhados por *E. invadens* e *T. grayi* e 10,15% (314/3.092) são grupos compartilhados por *C. Muris* e *T. Grayi* (Tabela 2). O OMA, por outro lado, inferiu 1.231 grupos ortólogos, cobrindo 10,72% do conjunto de dados (2.844/26.514 proteínas) e, neste caso, observou-se a seguinte distribuição: 3 espécies: 31,03% (382/1.231) e com 2 espécies, 30,78% (379/1.231) são grupos compartilhados por *E. invadens* e *T. grayi*, 21,68% (267/1.231) por *C. muris* e *T. grayi* e 16,49% (203/1.231) por *E. invadens* e *C. muris*. Uma síntese dos grupos homólogos inferidos nesse estudo é apresentada na Tabela 2.

**Tabela 2: Grupos homólogos inferidos pelo OrthoMCL e o OMA.**
Ortólogos do OrthoMCL: 1.481 (sendo 494 exclusivamente ortólogos) e
Ortólogos do OMA: 1.231;

| Grupos homólogos | Ortólogos | | | | Parálogos | | | |
|---|---|---|---|---|---|---|---|---|
| | *C. muris/ E. Invadens/ T. grayi* | *C. muris/ E. invadens* | *C. muris/ T. grayi* | *E. Invadens/ T. grayi* | *C. muris* | *E. invadens* | *T. grayi* | **Total** |
| **OrthoMCL** | 758 | 164 | 314 | 245 | 114 | 994 | 503 | 3.092 |
| **OMA** | 382 | 203 | 267 | 379 | N/A | N/A | N/A | 1.231 |

### 4.1.2 Análise dos resultados do OMA e OrthoMCL

Uma síntese das 3 categorias criadas pelo nível de concordância entre OrthoMCL e OMA resultante da abordagem de reconciliação proposta neste estudo, é mostrada na (Tabela 3).

**Tabela 3: Distribuição dos grupos homólogos em 3 categorias de acordo com o nível de concordância entre as ferramentas de inferência de homologia.**
Categoria I: “Grupos Idênticos", correspondem a 14,37% (445/3.092) dos grupos homólogos do OrthoMCL e a 36,14% (445/1.231) dos grupos homólogos de OMA, é formada por grupos homólogos onde os dois programas concordaram, contendo exatamente as mesmas proteínas; Categoria II: "Divergentes com interseções": correspondem 19,27% (596/3.092) dos grupos OrthoMCL e a 58,4% (719/1.231) dos grupos OMA, é formada por grupos homólogos diferentes entre os dois programas, mas com pelo menos uma proteína em comum; Categoria III: "Sem interseções": correspondem a 66,33% (2.051/3.092) dos grupos OrthoMCL e a 5,44% (67/1.231) dos grupos que o OMA inferiu e é formada por grupos homólogos onde nenhuma proteína é compartilhada entre os resultados dos dois programas.

| Categorias | Categoria I: Grupos Idênticos | Categoria II: Divergentes com interseções | Categoria III: Sem intersecções | Total |
|---|---|---|---|---|
| **OrthoMCL** | 445 | 596 | 2.051 | 3.092 |
| **OMA** | 445 | 719 | 67 | 1.231 |

### 4.1.3 Inferência de distância evolutiva para o OrthoMCL e OMA

Para verificar qual ferramenta de inferência de homologia seria capaz de inferir homólogos mais distantes, foi necessário normalizar os dados, visto que o OMA não inferiu parálogos, foram considerados apenas os grupos 494 grupos OrthoMCL exclusivamente ortólogos.

Os resultados do programa Belvu mostraram que o OMA, proporcionalmente, criou o maior número de grupos homólogos com uma distância evolutiva máxima de até 150 PAM: 84,57% (1.042/1.232), enquanto o OrthoMCL criou 66,59% (329/494) de seus grupos exclusivamente ortólogos, até esta distância. No entanto, para os grupos com distância evolutiva no intervalo entre 151 e 215,3 PAM (a maior distância evolutiva inferida nesta comparação), OrthoMCL inferiu mais grupos homólogos: 33,4% (165/494) versus 15,34% (189/1.232) inferido por OMA (Figura 4-2). Por outro lado, grupos homólogos com parálogos inferidos por OrthoMCL (que não estão nesta análise) atingiram distâncias evolutivas até 300 PAM.

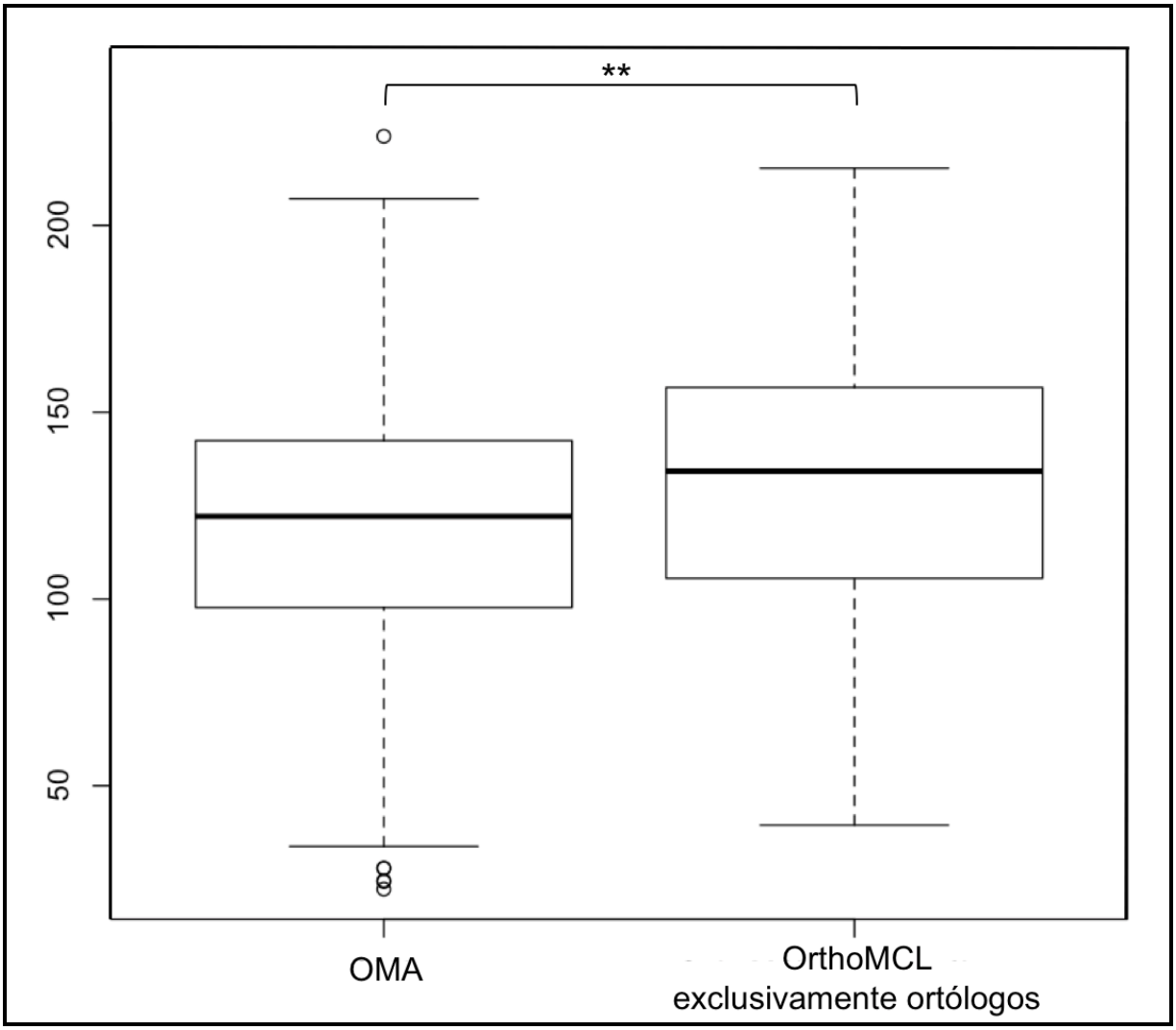



**Figura 4-2: Diagrama de caixa apresentando os valores das distâncias evolutivas dentro dos grupos OMA e dos grupos do OrthoMCL que possuem somente ortólogos.**

As distâncias não apresentaram distribuição normal no teste de normalidade Shapiro-Wilk (OrthoMCL – exclusivamente ortólogos p-value: 0.001175; OMA p-value: 0.00000007573). A figura mostra que os grupos do OrthoMCL apresentaram maiores distâncias estatisticamente significativas em relação aos do OMA: **p-value: 0.00000000787 no resultado do teste Wilcoxon-Mann-Whitney. As linhas espessas dentro dos retângulos representam as respectivas medianas, os lados inferiores e superiores dos retângulos representam os primeiros e terceiros quartis, enquanto que as barras horizontais situadas nas extremidades representam os limites superiores e inferiores das respectivas distâncias.

### 4.1.4 Identificação de domínios conservados (CDD)

Os resultados da análise RPS-BLAST mostraram que, dentre as proteínas de *E. invadens*, 59,13% (7.094/11.997) apresentam apenas um domínio conservado (CDD) e 2,65% (319/11.997) apresentam mais de um domínio conservado identificado. Em relação às proteínas de *C. muris*, 61,79% (2.431/3.934) mostram apenas 1 domínio conservado identificado e 2,87% (113/3.934) apresentam 2 ou mais domínios conservados. Já em *T. grayi,* 45,52% (4.818/10.583) apresentam um domínio conservado identificado e 2,41% (256/10.583) mais do que um domínio conservado (Figura 4-3).

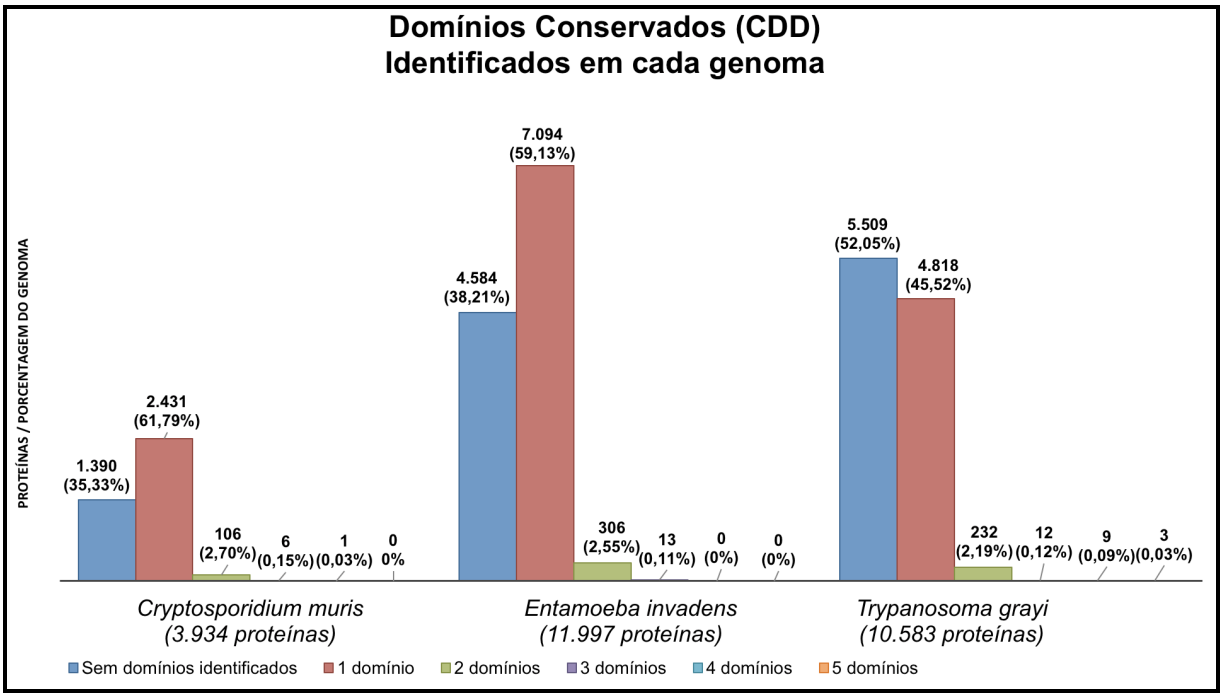



**Figura 4-3: Domínios conservados (CDD) identificados em cada genoma.**
A inferência de domínios conservados para cada proteína foi obtida usando o RPS-BLAST versão 2.2.13, contra a versão 3.12 da base de dados CDD com e-value de 1E-05.

### 4.1.5 Identificação de famílias de proteínas (Pfam-A)

Os resultados da ferramenta CLADE mostraram que *E. invadens* têm 49,14% (5.895/11.997) de suas proteínas pertencentes à apenas uma família de proteínas e 24,88% (2.985/11.997) pertencentes à mais de uma. Quanto às proteínas de *C. muris*, 50,66% (1.993/3.934) pertencem à apenas uma família de proteínas e 29,05% (1.143/3.934) a mais de uma família. E a análise de proteínas de *T. grayi* mostrou que 42,98% (4.549/10.583) tiveram uma família de proteínas identificada e 19,12% (2.024/10.583), mais de uma (Figura 4-4).

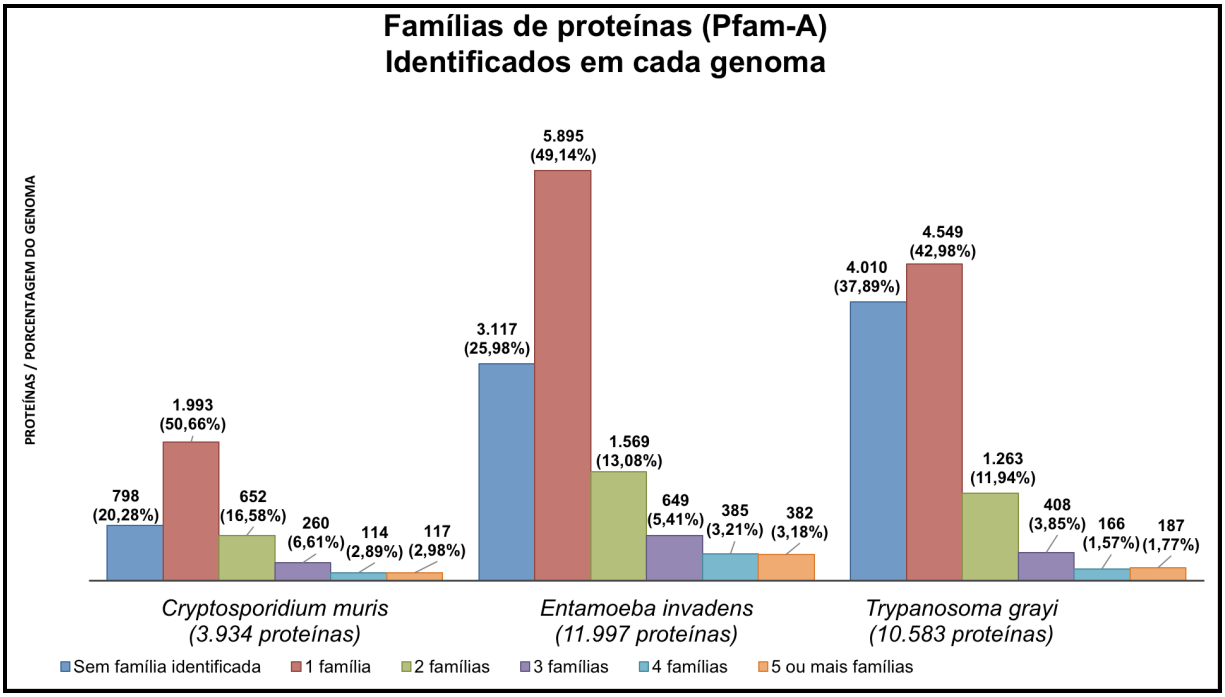



**Figura 4-4: Famílias de proteínas (Pfam-A) identificadas em cada genoma.**
A inferência de famílias de proteínas para cada proteína foi obtida com CLADE usando a base de dados do Pfam 27.0 com o e-value 1E-05.

## 4.2 Inferência de homologia baseada em uma abordagem de reconciliação para a genômica comparativa de protozoários

### 4.2.1 Resultados da validação dos homólogos da Categoria II: "Divergentes com as interseções" – Inferência de Supergrupos

Com base nos grupos homólogos da categoria II "Divergentes com interseções" (Capítulo 4.1), foram inferidos 537 grupos homólogos resultantes da reconciliação entre os resultados OMA e OrthoMCL, que são novos grupos concatenados. Destes, 76,16% (409/537) são grupos que não aumentaram o número de proteínas em relação a todos os grupos homólogos que o originaram e, portanto, grupos OrthoMCL com um ou mais grupos OMA contidos ou vice-versa. Esses 409 grupos concatenados foram então descartados da nossa análise.

Além disso, 23,83% (128/537) dos grupos inferidos com o algoritmo de reconciliação tiveram aumento no número de proteínas (em relação a todos os grupos homólogos que a originaram), esses grupos são grupos concatenados aumentados. Entre estes grupos, 46,09% (59/128) foram validados, apresentando o mesmo domínio conservado (CDD) ou pertencendo a mesma família de proteínas (Pfam-A) em todas as suas proteínas e serão chamados a partir daqui de Supergrupos homólogos. Entre os grupos concatenados aumentados não validados, 63,76% (44/69) apresentaram pelo menos uma proteína com domínio conservado diferente e com família de proteínas (Pfam-A) diferente, 36,23% (25/69) apresentaram pelo menos 1 proteína sem domínio conservado e família de proteínas identificada (Figura 4-5).

Os 59 supergrupos apresentaram a seguinte distribuição: 59,32% (35/59) compartilhados pelas 3 espécies e foram formados por 46 grupos OrthoMCL e 67 grupos OMA, 1,69% (1/59) compartilhado por *C. muris* e *E. invadens* e ele foi formado por 1 grupo OrthoMCL e 1 grupo OMA, 10,16% (6/59) compartilhado por *C. muris* e *T. grayi* e foram formados por 7 grupos OrthoMCL e 6 grupos OMA e 28,81% (17/59) compartilhados por *T. grayi* e *E. invadens* e foram formados por 21 grupos OrthoMCL e 20 grupos OMA.

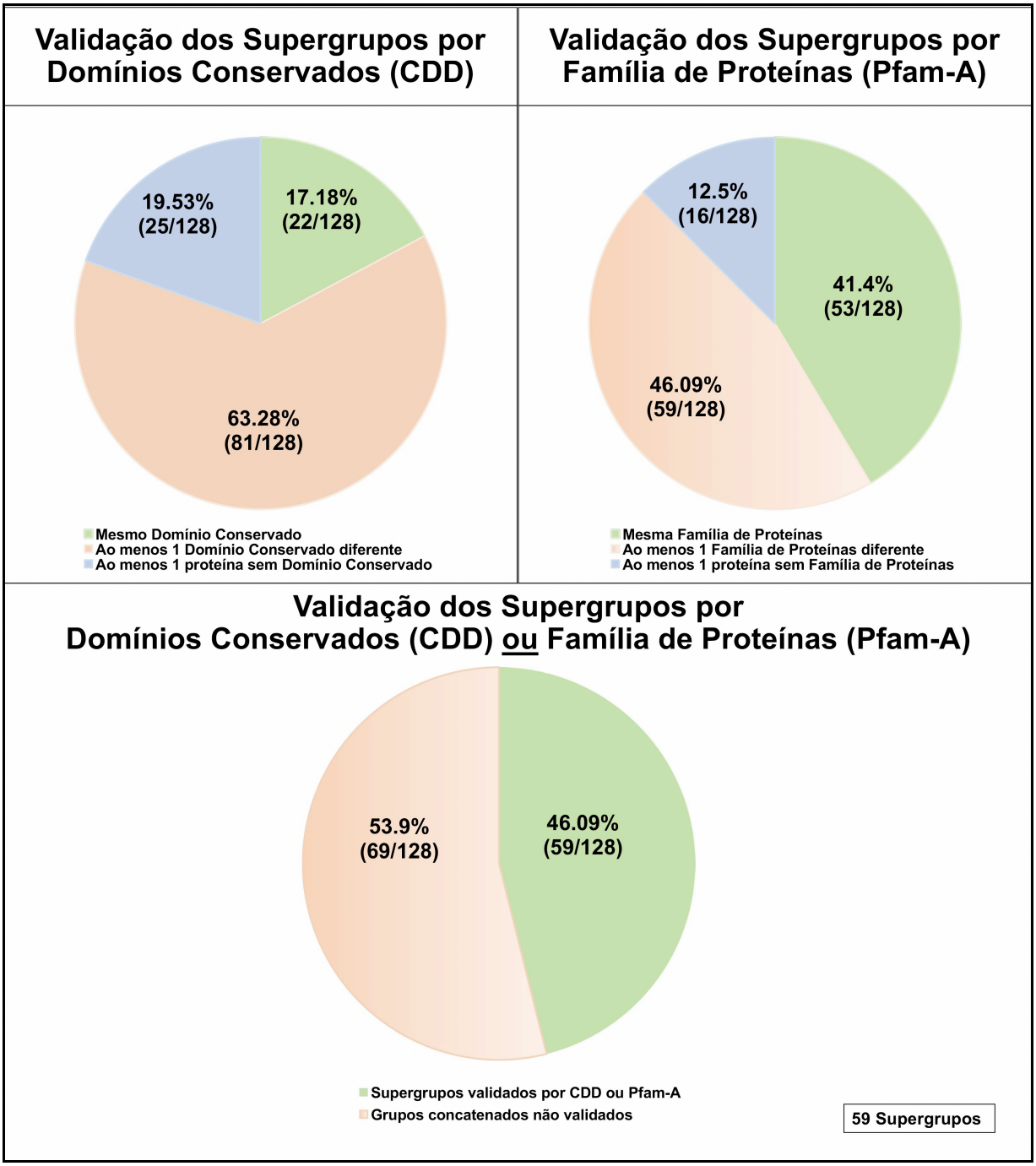



**Figura 4-5: Validação dos Supergrupos por domínio conservado (CDD) e família de proteínas (Pfam-A).**
Os Supergrupos homólogos foram validados verificando se todas as suas proteínas possuem o mesmo domínio conservado (CDD) (abordagem 1) ou pertencem à mesma família de proteínas (Pfam-A) (abordagem 2). 46,09% (59/128) apresentaram exatamente o mesmo domínio conservado ou a mesma família de proteínas em suas proteínas.

Após a inferência dos supergrupos homólogos, houve uma diminuição no número de grupos homólogos inferidos por OrthoMCL e OMA, já que alguns de seus grupos foram fundidos para formar os Supergrupos. O total de homólogos inferidos neste estudo é a soma dos homólogos inferidos por OrthoMCL com os homólogos inferidos pela OMA e os 59 Supergrupos homólogos. A nova distribuição dos grupos homólogos inferidos neste estudo é mostrada na Figura 4-6.

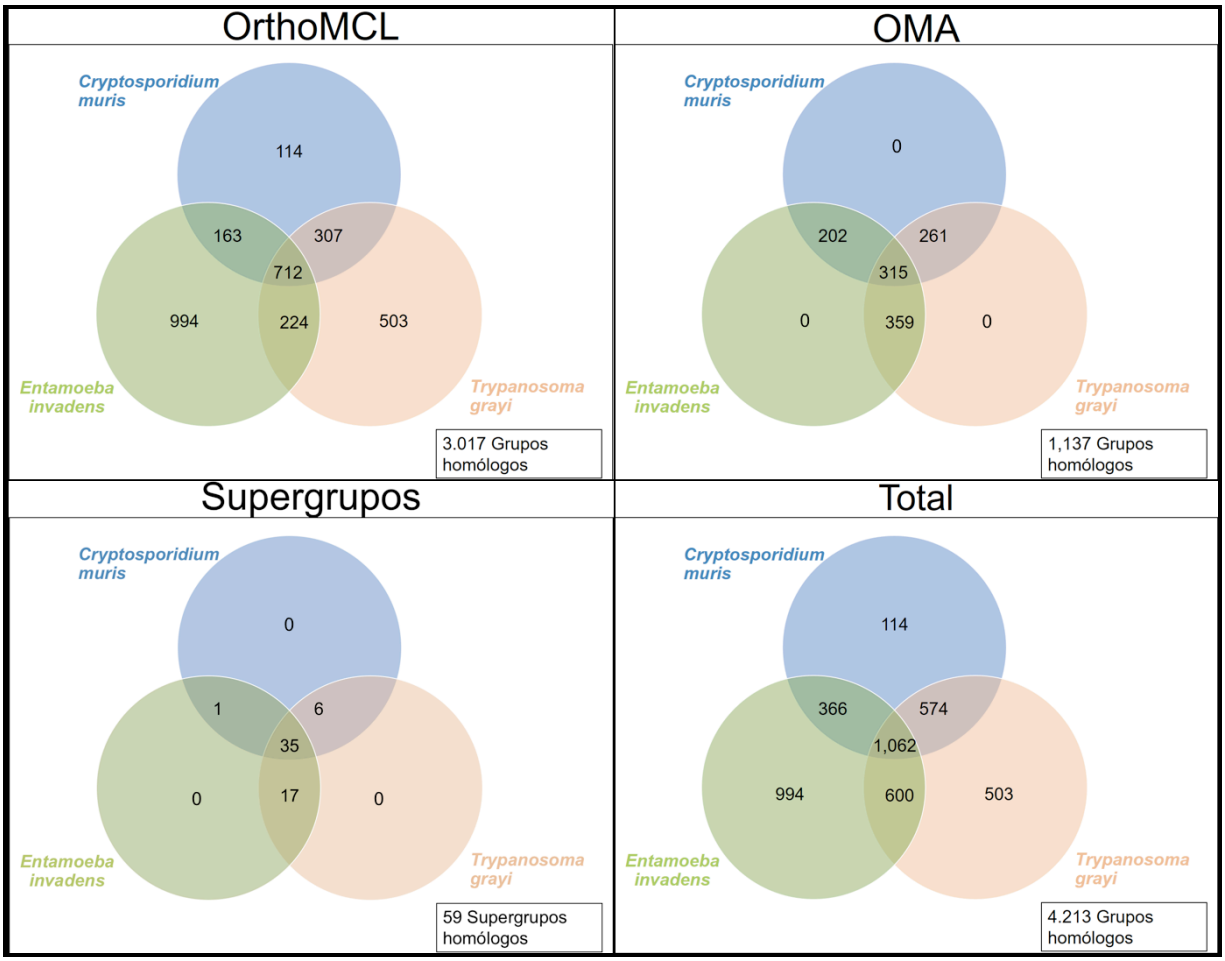



**Figura 4-6: Nova distribuição dos grupos homólogos inferidos neste estudo.**

### 4.2.2 Identificação de domínios conservados (CDD) nos Supergrupos

Os resultados da análise RPS-BLAST mostraram que nos Supergrupos homólogos validados por CDD, todas as proteínas de *E. invadens* 100% (55/55), apresentam apenas um domínio conservado (CDD). Em relação às proteínas de *C. muris* em Supergrupos validados por CDD, 95,91% (23/24) mostram apenas 1 domínio conservado identificado e 4,16% (1/24) apresentam 2 domínios conservados. Já em *T. grayi,* 97,29% (36/37) apresentam um domínio conservado identificado e 2,7% (1/37) mais do que um domínio conservado (Figura 4-7).

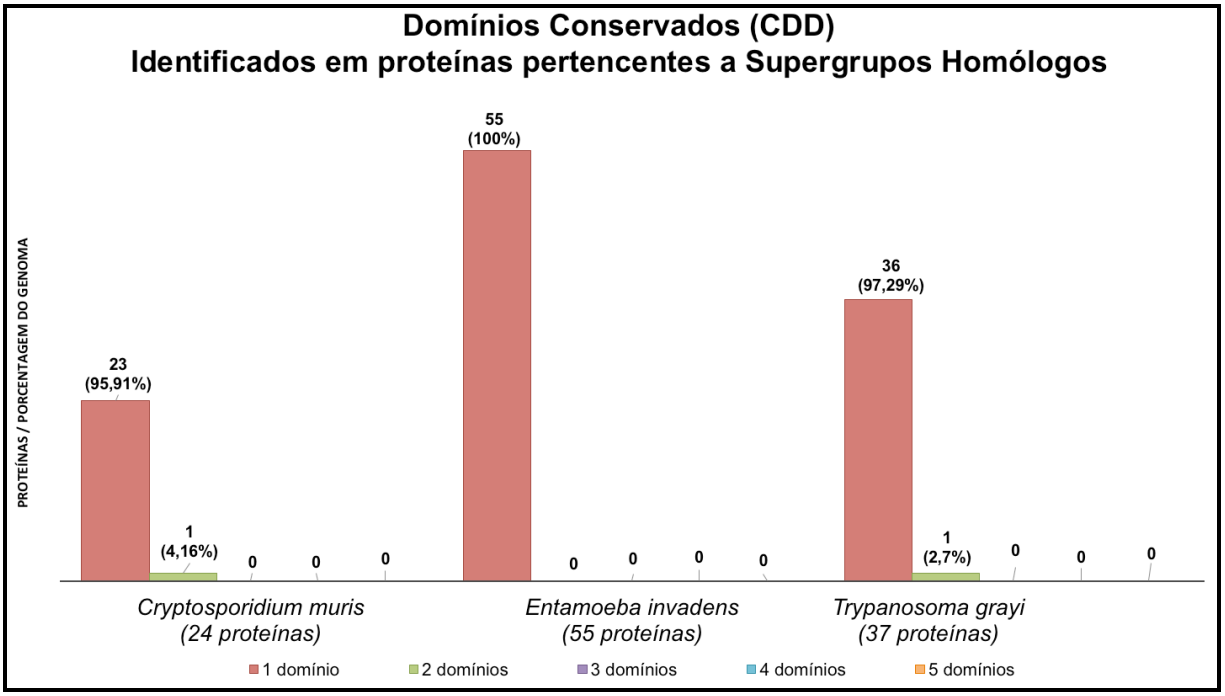



**Figura 4-7: Distribuição das 116 proteínas presentes nos 22 Supergrupos homólogos validados por CDD.**

### 4.2.3 Identificação de famílias de proteínas (Pfam-A) nos Supergrupos

Os resultados do CLADE mostraram que nos Supergrupos homólogos validados por Pfam-A, *E. invadens* têm 94,11% (128/136) de suas proteínas pertencentes à apenas uma família de proteínas e 5,88% (8/136) pertencentes à mais de uma. Quanto às proteínas de *C. muris*, 85,45% (47/55) pertencem à apenas uma família de proteínas e 14,54% (8/55) a mais de uma família. E a análise de proteínas de *T. grayi* mostrou que 83,2% (104/125) tiveram uma família de proteínas identificada e 16,8% (21/125), mais de uma (Figura 4-8).

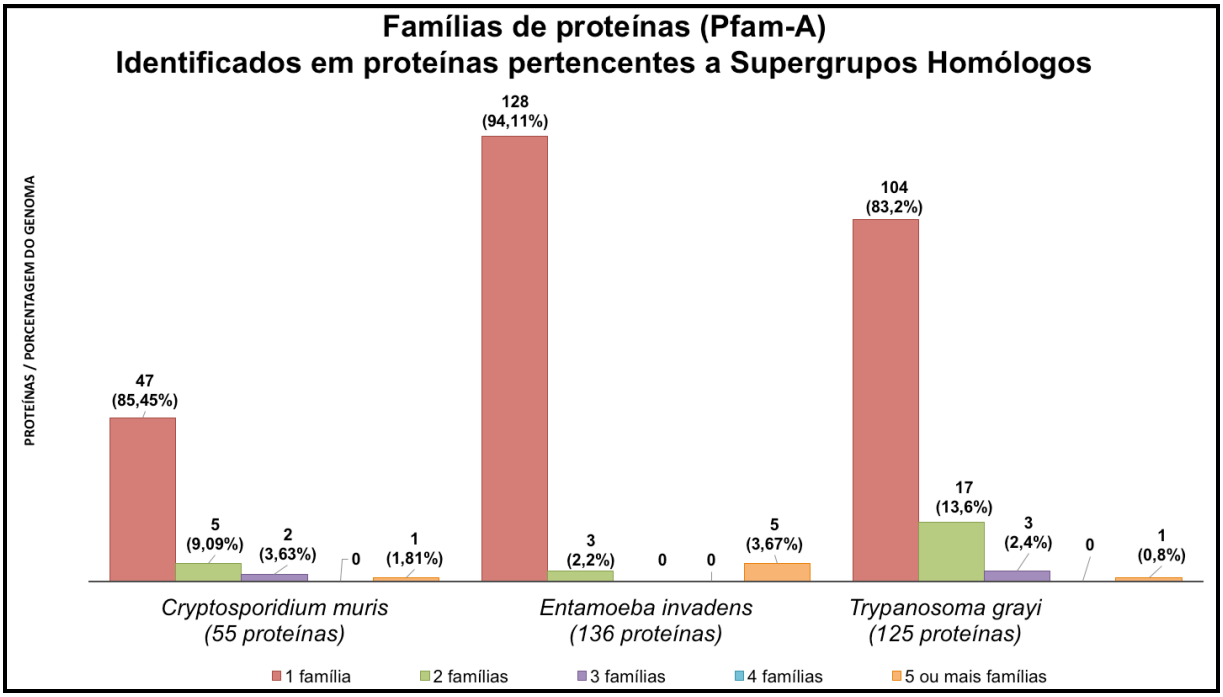



**Figura 4-8: Distribuição das 315 proteínas presentes nos 53 Supergrupos homólogos validados por Pfam-A.**

### 4.2.4 Comparação com as bases de dados SUPERFAMILY e Pfam Clans

A comparação dos 59 Supergrupos homólogos com as bases de dados SUPERFAMILY e Pfam Clans mostrou que 81,35% (48/59) dos Supergrupos possuem todas as suas proteínas pertencentes à mesma Superfamília. No entanto, embora 89,83% (53/59) dos Supergrupos tenham todas as proteínas pertencentes à mesma família de proteínas (Pfam-A), apenas 77,96% (46/59) possuem proteínas pertencentes ao mesmo Pfam clan. Isso se deve ao fato de que as famílias de proteínas (Pfam-A) identificadas em 20,33% (12/59) dos Supergrupos não pertencem a nenhum Pfam clan.

### 4.2.5 Análise da conservação das sequências dos Supergrupos

Foram criados alinhamentos múltiplos para todos os 59 Supergrupos e o seus tamanhos variaram entre 150 a 1.311 aminoácidos. O alinhamento múltiplo apresentou mais de 34,51% de identidade média entre sequências de cada Supergrupo e 32,29% de identidade média entre as duas sequências mais distantes.

Como parâmetro de comparação, os grupos concatenados que não foram validados (69/128) por CDD ou Pfam, apresentaram alinhamentos múltiplos com tamanho variando de 129 a 19.451 aminoácidos com 29,04% de média de identidade entre suas sequências e 27,71% de média entre as suas sequências mais distantes.

### 4.2.6 Filogenia

Árvores filogenéticas baseadas em Máxima verossimilhança foram inferidas para todos os Supergrupos homólogos com 4 ou mais proteínas, equivalente a 81,35% (48/59) do total.

### 4.2.7 Inferência da distância evolutiva para os Supergrupos

A análise dos resultados do programa Belvu mostrou que os Supergrupos homólogos inferidos neste estudo apresentaram maiores distâncias evolutivas quando comparados aos grupos homólogos inferidos por OrthoMCL e OMA (Figura 4-9). Em relação aos Supergrupos, a distância evolutiva média foi de 151,77 PAM, com uma distância evolutiva mínima de 80,37 PAM e distância evolutiva máxima de 292,28 PAM.

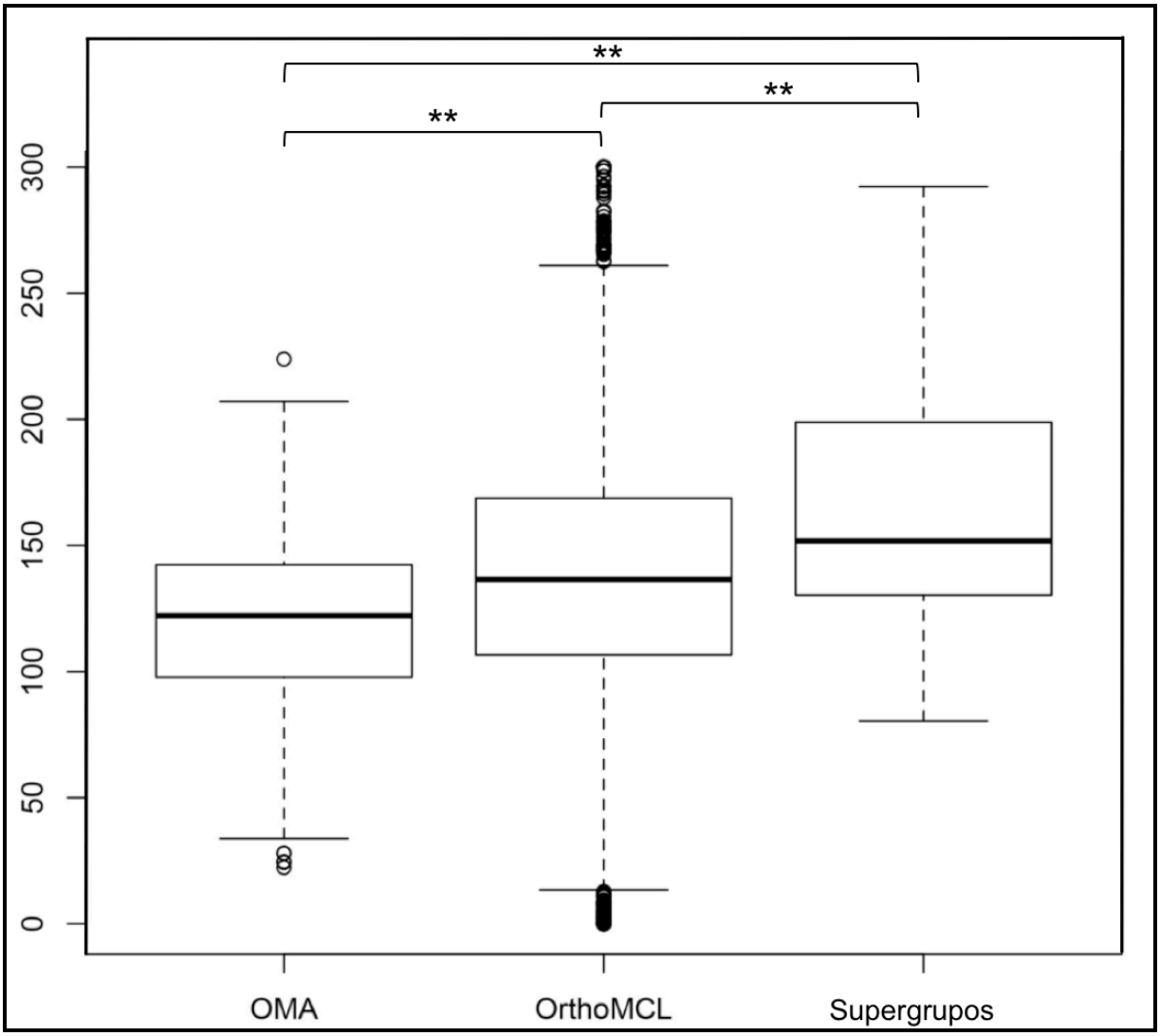



**Figura 4-9: Diagrama de caixa representando as distâncias evolutivas dos Supergrupos homólogos e dos grupos OrthoMCL e OMA originais.**

As distâncias não apresentaram distribuição normal no teste de normalidade Shapiro-Wilk (OrthoMCL p-value: 2.2e-16; OMA p-value: 0.00000007573; Supergrupos p-value: 0.006784). A figura mostra que os Supergrupos homólogos apresentaram maiores distâncias estatisticamente significativas em relação aos do OrthoMCL: **p-value: 0.0005238 e do OMA: **p-value: 0.0000004886 no resultado do teste Wilcoxon-Mann-Whitney. As linhas espessas dentro dos retângulos representam as respectivas medianas, os lados inferiores e superiores dos retângulos representam os primeiros e terceiros quartis, enquanto que as barras horizontais situadas nas extremidades representam os limites superiores e inferiores das respectivas distâncias.

### 4.2.8 Categorização funcional dos Supergrupos

A categorização funcional dos 59 Supergrupos, mostrou que 32,2% (19/59) pertence a uma categoria funcional S com função desconhecida ou “function unknown”, 3,38% (2/59) ainda não tiveram categoria funcional inferida, 22,03% (13/59) pertencem a Categoria funcional T: " Signal transduction mechanisms", 6,77% (4/59) pertence à categoria funcional O: " Posttranslational modification, protein turnover, chaperones", 10,16% (6/59) pertencem à categoria funcional E: “Amino acid transport and metabolism”, 5,08% (3/59) pertencem à categoria funcional J: “Translation, ribosomal structure and biogenesis”, 6,77% (4/59) pertencem à categoria funcional U: “Intracellular trafficking, secretion, and vesicular transport”. As categorias funcionais B: “Chromatin structure and dynamics”, L: “Replication”, P: “Inorganic ion transport and metabolism”, C: “Energy production and conversion”, H: “Coenzyme transport and metabolism” and K: “Transcription” representaram 1,69% (1/59) cada, com um Supergrupo pertencente a cada uma dessas categorias. Além disso, 2 Supergrupos apresentaram proteínas que pertencem a categorias funcionais distintas, da seguinte forma: 1,69% (1/59) pertence às categorias J e O e 1,69% (1/59) pertence às categorias P e U (Figura 4-10). Portanto, 96,61% (57/59) dos Supergrupos apresentaram todas as suas proteínas pertencentes à mesma categoria funcional.

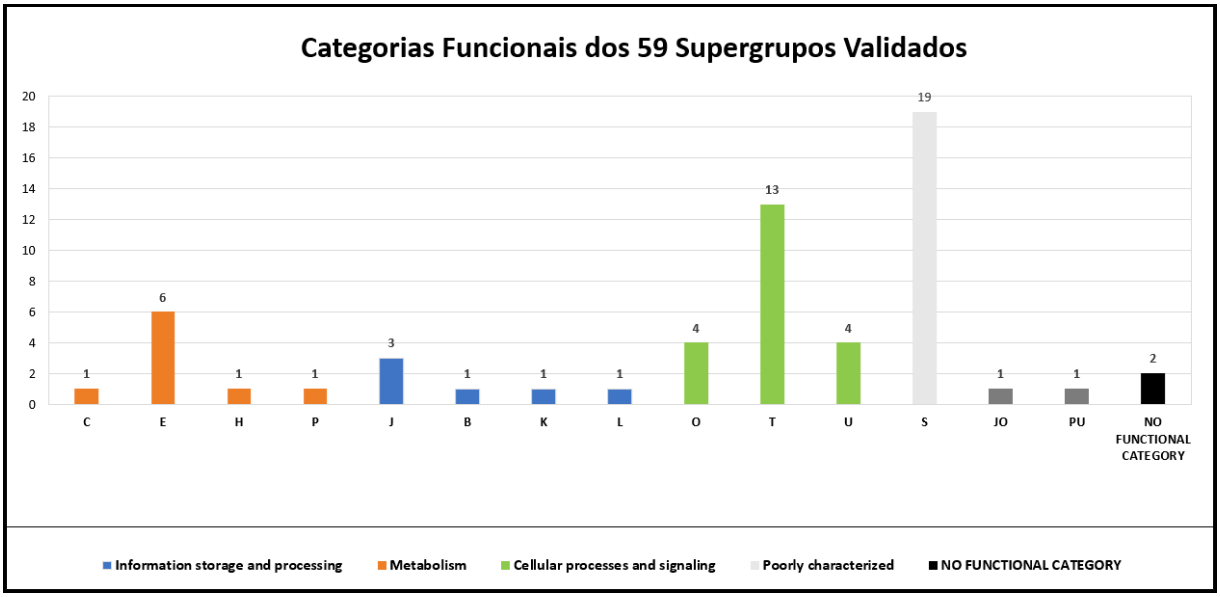



**Figura 4-10: Categorias funcionais inferidas para as proteínas dos 59 Supergrupos.**
[B] Chromatin structure and dynamics; [C] Energy production and conversion; [E] Amino acid transport and metabolism; [H] Coenzyme transport and metabolism; [J] Translation, ribosomal structure and biogenesis; [K] Transcription; [L] Replication; [O] Posttranslational modification, protein turnover, chaperones; [P] Inorganic ion transport and metabolism; [S] Function unknown; [T] Signal transduction mechanisms; [U] Intracellular trafficking, secretion, and vesicular transport.

### 4.2.9 Vias Metabólicas do KEGG dos Supergrupos

O resultado da análise realizada pelo BLASTP entre a base de dados de genes eucariotas e procariotas do KEGG mostrou que 61,01% (36/59) dos Supergrupos possuem proteínas que participam de pelo menos uma via KEGG. Desses, 16,66% (6/36) são pertencentes à via "Metabolismo", e escolhemos estes para serem utilizados como estudo de caso, (Tabela 4).

**Tabela 4: Supergrupos escolhidos para serem usados como caso de estudo por pertencerem à via "metabolismo".**

| | Supergrupo | Nº de proteínas | Best hit KEGG | Organismos |
|---|---|---|---|---|
| (a) | SG_1364 | 4 | K01507 | *C. muris* / |
| (b) | SG_1363 | 4 | K19787 | *E. invadens* / |
| (c) | SG_1634 | 4 | K04487 | |
| (d) | SG_843 | 5 | K10251 | *T. grayi* |
| (e) | SG_1241 | 5 | K01697 | |
| | | | K01738 | *E. invadens* / *T. grayi* |
| (f) | SG_711 | 5 | K01760 | |

## 4.3 Inferência e análise de homólogos distantes em Protozoa por meio da comparação pHMM - pHMM (perfis de Modelo Oculto de Markov) visando a identificação de superfamílias

### 4.3.1 Comparação pHMM - pHMM com o COMA

O resultado da comparação pHMM - pHMM pela ferramenta COMA entre os grupos homólogos previamente inferidos pelo OMA e pelo OrthoMCL mostrou qual o melhor hit para cada um desses grupos e o nosso script apontou quais deles tiveram melhor hit recíproco com outro grupo.

Primeiramente, foram identificados hits recíprocos entre os grupos ortólogos e parálogos inferidos pelo OrthoMCL separadamente. E depois, o mesmo foi feito entre os grupos ortólogos inferidos pelo OMA.

Entre os 1.611 grupos parálogos inferidos pelo OrthoMCL, foram identificados 48 grupos com hit recíproco, que quando unidos (cada grupo com seu melhor hit) formaram 24 grupos homólogos distantes prováveis, ainda pendentes de validação na base de dados SUPERFAMILY.

Já entre os 1.481 grupos ortólogos inferidos pelo OrthoMCL, foram identificados 148 grupos com hit recíproco, totalizando 74 grupos homólogos distantes prováveis.

E entre os 1.232 grupos ortólogos inferidos pelo OMA, foram identificados 256 com hit recíproco, totalizando 128 grupos homólogos distantes prováveis.

### 4.3.2 Validação dos novos homólogos distantes inferidos pelo COMA

Os grupos homólogos distantes prováveis formados pela junção de grupos homólogos com melhor hit recíproco pHMM – pHMM (COMA) tiveram como método de validação a identificação de domínios pertencentes à mesma superfamília (análise feita utilizando a base de dados SUPERFAMILY) nos dois grupos homólogos que os originaram (grupos com melhor hit recíproco). Depois de validados usando a base de dados SUPERFAMILY, estes potenciais homólogos distantes serão os novos homólogos distantes OrthoMCL/COMA e OMA/COMA. Os que não forem validados, continuarão sendo potenciais homólogos distantes.

Os resultados mostraram que 40% (10/25) dos homólogos distantes gerados a partir de grupos parálogos do OrthoMCL, 45,83% (11/24) foram validados, apresentando proteínas associadas a mesma superfamília em grupos com melhores hits recíprocos (Figura 4-11), seguindo o mesmo padrão, 87,83% (65/74) dos homólogos distantes prováveis gerados a partir de grupos ortólogos do OrthoMCL foram validados (Figura 4-12) e o mesmo pode ser dito sobre 89,06% (114/128) dos homólogos distantes formados a partir dos grupos ortólogos inferidos pelo OMA (Figura 4-13).

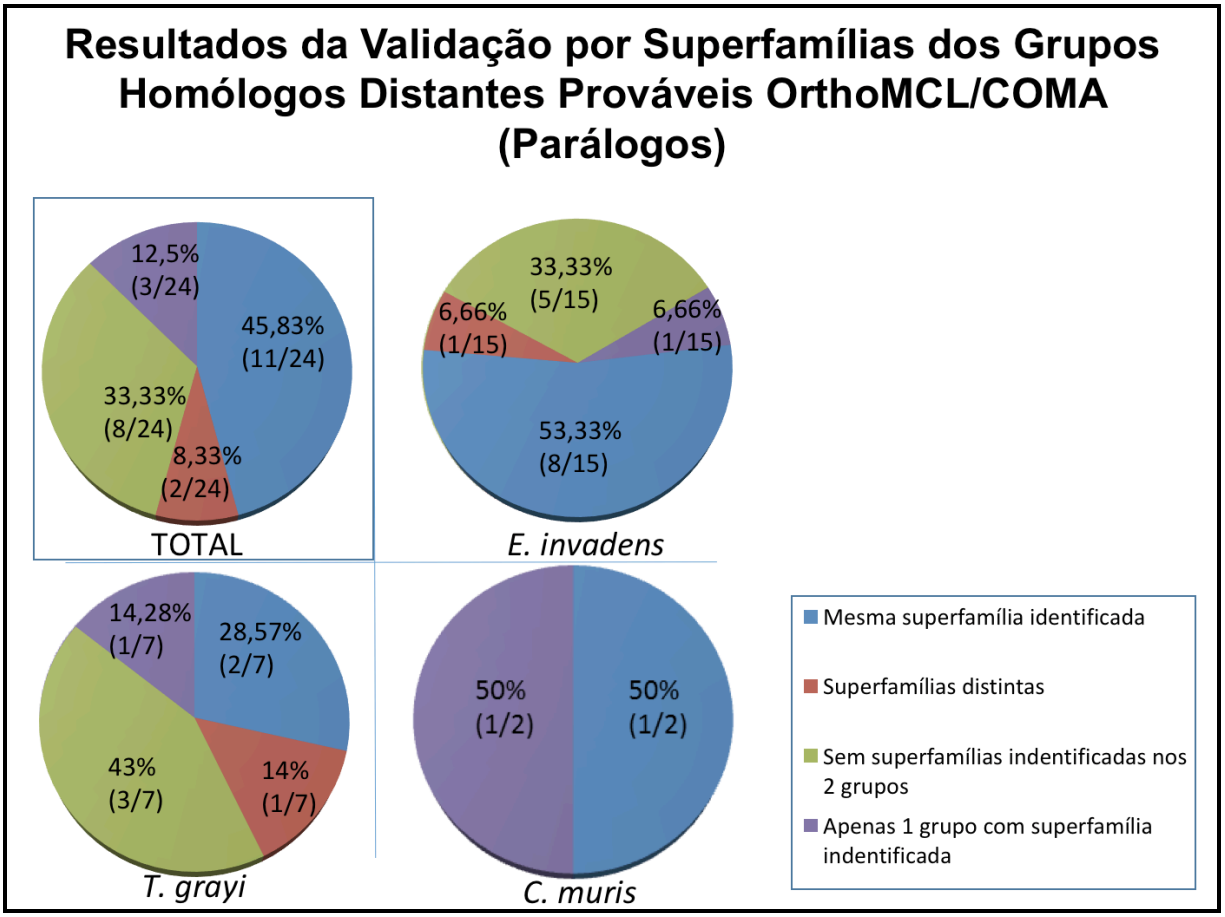



**Figura 4-11: Resultado da validação dos grupos homólogos distantes prováveis identificados entre melhores hits recíprocos entre os grupos parálogos inferidos pelo OrthoMCL, utilizando a base de dados SUPERFAMILY.**

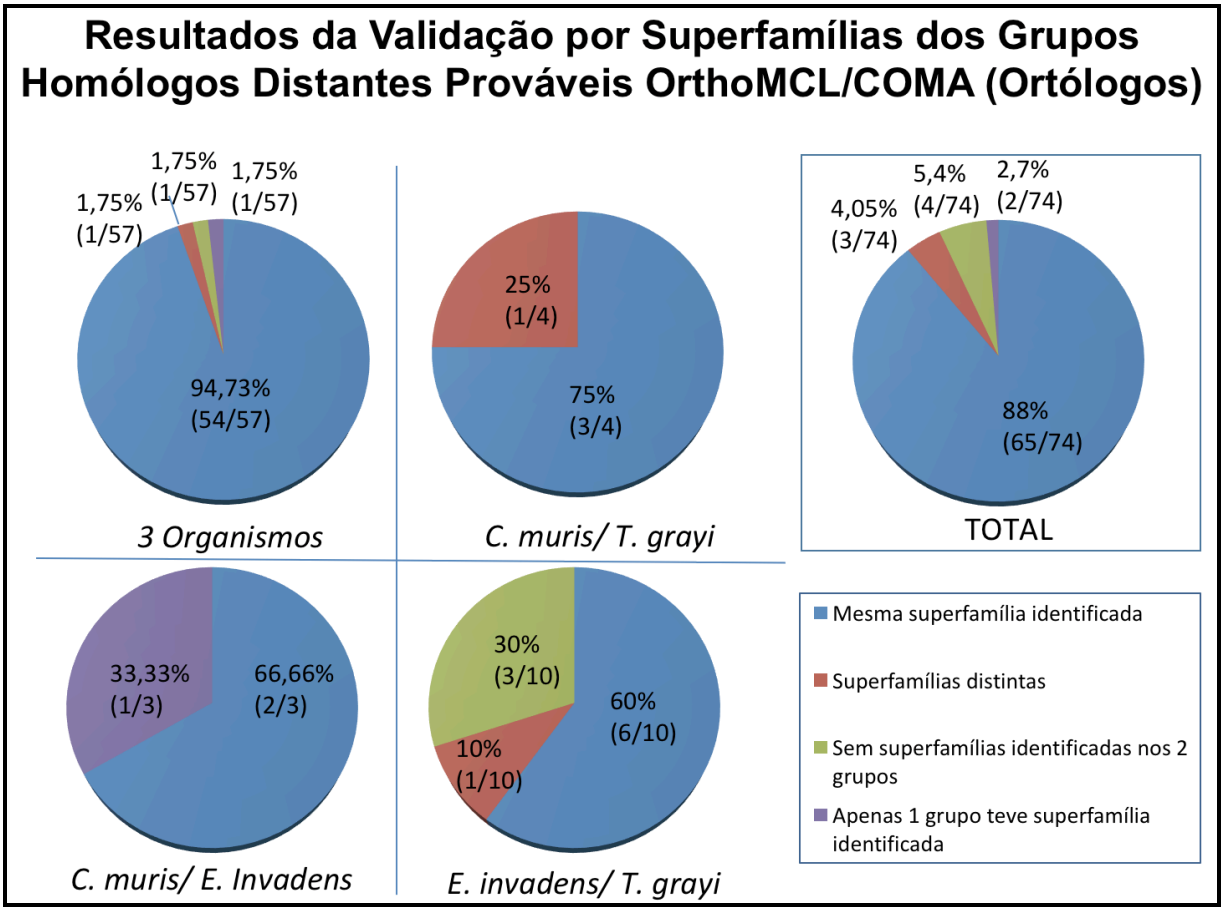



**Figura 4-12: Resultado da validação dos grupos homólogos distantes prováveis identificados entre melhores hits recíprocos entre os grupos ortólogos inferidos pelo OrthoMCL, utilizando a base de dados SUPERFAMILY.**

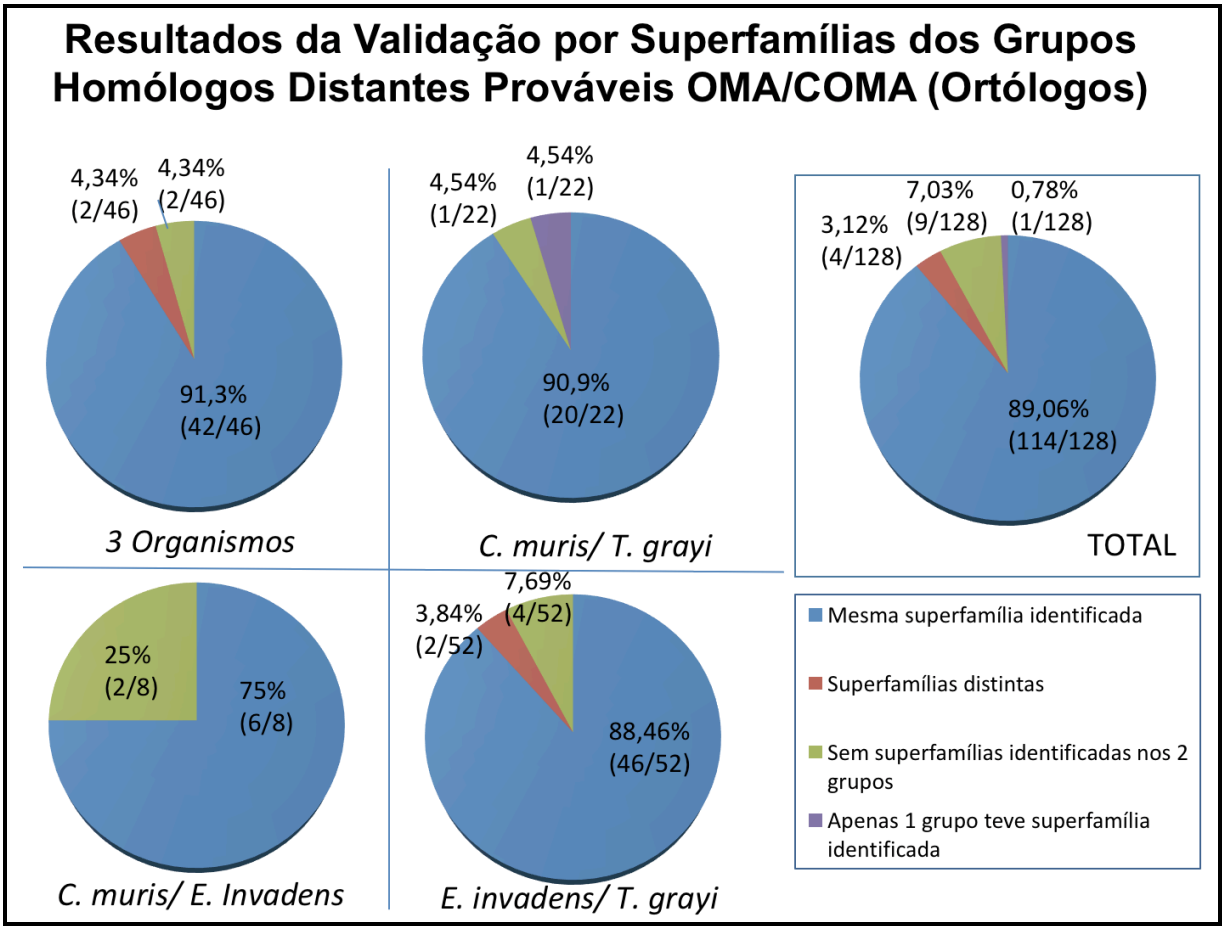



**Figura 4-13: Resultado da validação dos grupos homólogos distantes prováveis identificados entre melhores hits recíprocos pHMM – pHMM entre os grupos ortólogos inferidos pelo OMA, utilizando a base de dados SUPERFAMILY.**

Como mostram os resultados, 6 homólogos distantes prováveis não puderam ser validados, pois tiveram superfamília identificada em apenas 1 dos 2 grupos homólogos com melhor hit recíproco pHMM - pHMM, indicando a possibilidade de transferência de anotação (apenas 1 grupo teve superfamília identificada, nas figuras Figura 4-11, Figura 4-12 eFigura 4-13). Entre os homólogos distantes prováveis OrthoMCL/COMA, podemos associar o grupo homólogo ORTHOMCL1733 (3 organismos) com a superfamília "Class II aaRS ABD-related", o ORTHOMCL2704 (*C.muris* e *E. invadens*) com a superfamília "ARM repeat", o ORTHOMCL2828 (*C.muris*) com a superfamília "PH domain-like", o ORTHOMCL244 (*E. invadens*) com a superfamília "GTPase activation domain, GAP", ORTHOMCL952 (*T. grayi*) com a superfamília "N-terminal nucleophile aminohydrolases (Ntn hydrolases)" e o OMA01176 (*C.muris* e *T. grayi*) com a superfamília "P-loop containing nucleoside triphosphate hydrolases" (Tabela 5).

**Tabela 5**: **Grupos OrthoMCL/OMA alvos de reanotação funcional**
Grupos sem anotação SUPERFAMILY com melhor hit recíproco pHMM – pHMM com grupo homólogo com superfamília associada na base de dados SUPERFAMILY.

| **Grupo OrthoMCL/OMA alvo de reanotação funcional** | **Superfamília a ser associada** |
|---|---|
| ORTHOMCL1733 | Class II aaRS ABD-related |
| ORTHOMCL2704 | ARM repeat |
| ORTHOMCL2828 | PH domain-like |
| ORTHOMCL244 | GTPase activation domain, GAP |
| ORTHOMCL952 | N-terminal nucleophile aminohydrolases |
| OMA01176 | P-loop containing nucleoside triphosphate hydrolases |

### 4.3.3 Criação de grafos para análises adicionais

Visando compreender melhor a dinâmica entre os homólogos distantes prováveis, foi gerado um grafo mostrando as relações entre até os 3 primeiros hits pHMM - pHMM dos 144 grupos ortólogos com os 3 organismos inferidos pelo OrthoMCL que tiveram melhor hit recíproco pHMM – pHMM. (Figura 4-14). Seguindo o mesmo padrão, também foi gerado um grafo mostrando as relações entre até os 3 primeiros hits pHMM - pHMM dos 92 grupos ortólogos com os 3 organismos inferidos pelo OMA que tiveram melhor hit recíproco pHMM – pHMM (Figura 4-15).

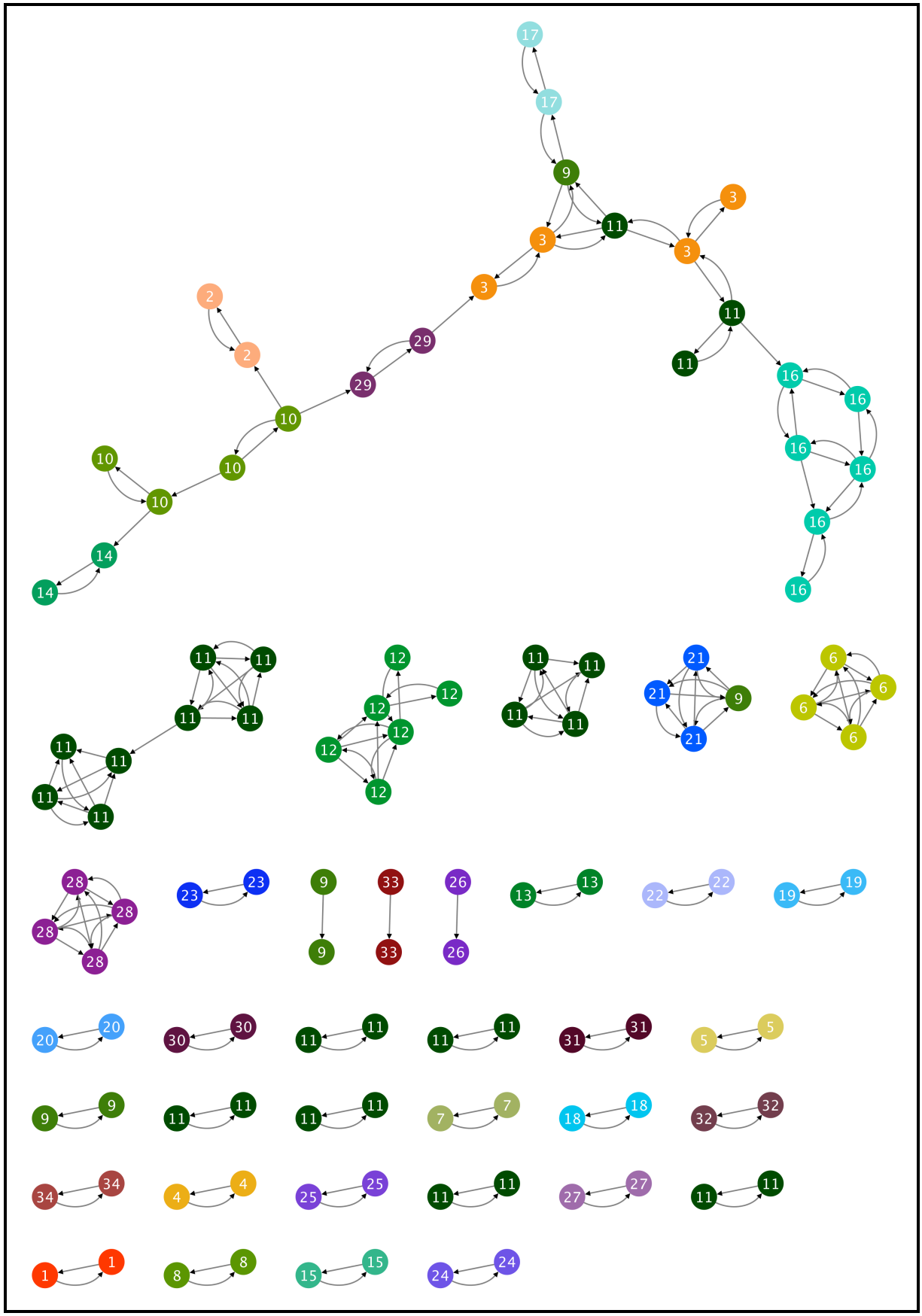

17
17
9
3
11
3
3
3
11
11
2
2
29
29
10
10
10
10
14
14
16
16
16
16
16
16
11
11
11
11
11
11
11
11
12
12
12
12
12
12
11
11
11
21
21
9
21
6
6
6
6
28
28
28
28
23
23
9
9
33
33
26
26
13
13
22
22
19
19
20
20
30
30
11
11
11
11
31
31
5
5
9
9
11
11
11
11
7
7
18
18
32
32
34
34
4
4
25
25
11
11
27
27
11
11
1
1
8
8
15
15
24
24

**Figura 4-14: Grafo representando as relações (arestas) entre os 114 grupos do OrthoMCL com ortólogos dos 3 organismos (nós) que tiveram melhor hit recíproco pHMM – pHMM. A figura mostra, quando possível, as relações entre os 3 melhores hits pHMM – pHMM e a superfamília identificada (código numérico e cor) em cada grupo.**

[1] Alpha-L RNA-binding motif; [2] ARM repeat; [3] beta and beta-prime subunits of DNA dependent RNA-polymerase; [4] EF-hand; [5] F1F0 ATP synthase subunit C; [6] GroEL equatorial domain-like; [7] Insert subdomain of RNA polymerase alpha subunit; [8] Metallo-dependent phosphatases; [9] No Superfamily; [10] Nucleotidylyl transferase; [11] P-loop containing nucleoside triphosphate hydrolases; [12] Protein kinase-like (PK-like); [13] S-adenosyl-L-methionine-dependent methyltransferases; [14] Tubulin nucleotide-binding domain-like; [15] UBC-like; [16] WD40 repeat-like; [17] Actin-like ATPase domain; [18] Acyl-CoA N-acyltransferases (Nat); [19] ArfGap/RecO-like zinc finger; [20] Chaperone J-domain; [21] Class II aaRS ABD-related; [22] Cyclophilin-like; [23] Cysteine proteinases; [24] Glucocorticoid receptor-like (DNA-binding domain); [25] Nop domain; [26] N-terminal nucleophile aminohydrolases (Ntn hydrolases); [27] PIN domain-like; [28] Nucleic acid-binding proteins; [29] Prokaryotic type I DNA topoisomerase; [30] Ribosomal protein S5 domain 2-like; [31] RNA-binding domain, RBD; [32] Terpenoid cyclases/Protein prenyltransferases; [33] Ypt/Rab-GAP domain of gyp1p; [34] Zinc beta-ribbon;

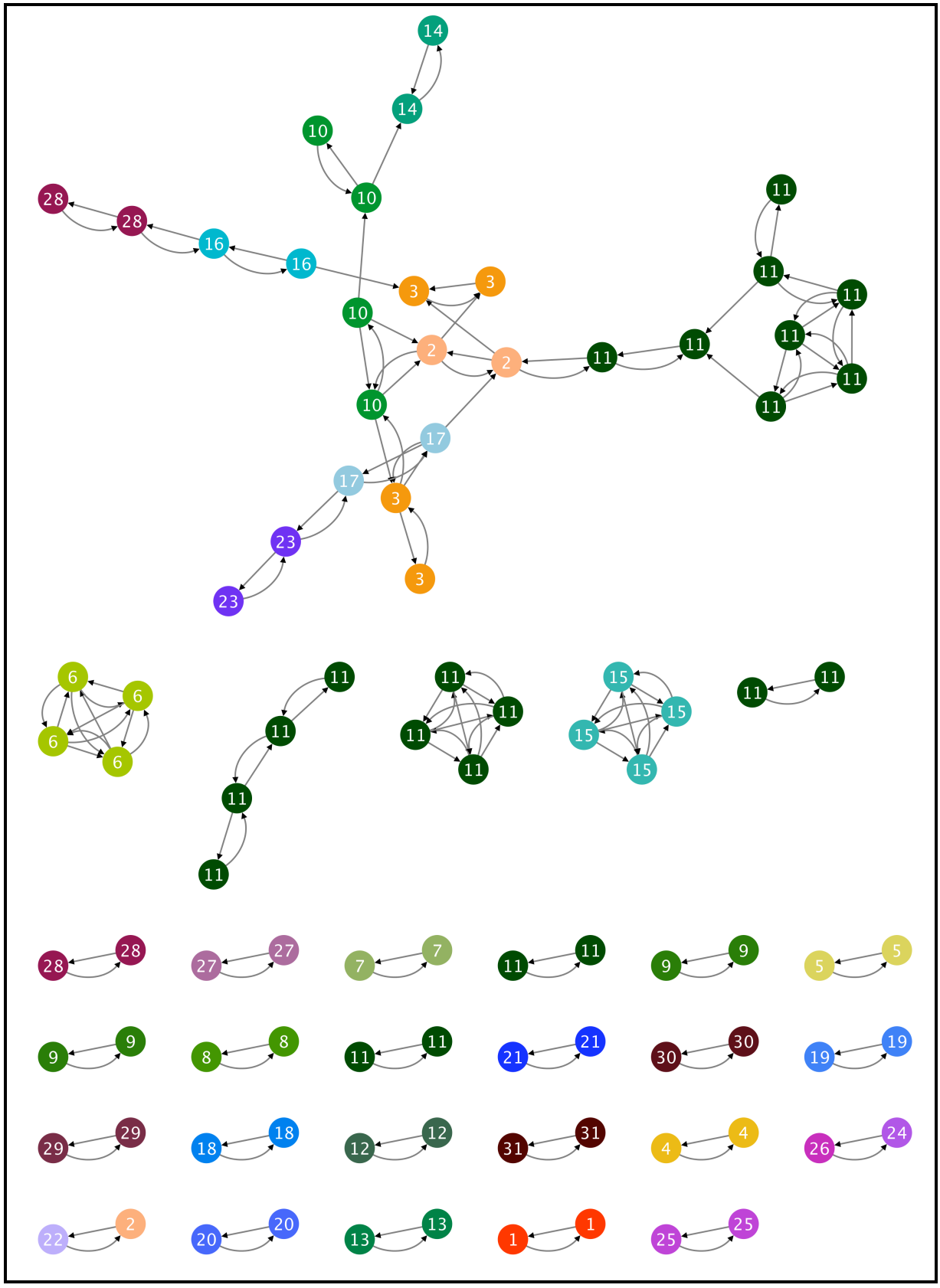



**Figura 4-15: Grafo representando as relações (arestas) entre os 92 grupos do OMA com ortólogos dos 3 organismos (nós) que tiveram melhor hit recíproco pHMM – pHMM. A figura mostra, quando possível, as relações entre os 3 melhores hits pHMM – pHMM e a superfamília identificada (código numérico e cor) em cada grupo.**

[1] Alpha-L RNA-binding motif; [2] ARM repeat; [3] beta and beta-prime subunits of DNA dependent RNA-polymerase; [4] EF-hand; [5] F1F0 ATP synthase subunit C; [6] GroEL equatorial domain-like; [7] Insert subdomain of RNA polymerase alpha subunit; [8] Metallo-dependent phosphatases; [9] No Superfamily; [10] Nucleotidylyl transferase; [11] P-loop containing nucleoside triphosphate hydrolases; [12] Protein kinase-like (PK-like); [13] S-adenosyl-L-methionine-dependent methyltransferases; [14] Tubulin nucleotide-binding domain-like; [15] UBC-like; [16] WD40 repeat-like; [17] Activating enzymes of the ubiquitin-like proteins; [18] Calcium ATPase, transmembrane domain M; [19] Class II aaRS and biotin synthetases; [20] Eukaryotic type KH-domain (KH-domain type I); [21] Histone-fold; [22] JAB1/MPN domain; [23] NAD(P)-binding Rossmann-fold domains; [24] PTPA-like; [25] Ribosomal protein L10-like; [26] Ribosomal protein L4; [27] SNARE-like; [28] Thioredoxin-like; [29] t-snare proteins; [30] Ubiquitin-like; [31] UDP-Glycosyltransferase/glycogen phosphorylase;

A lista de Superfamílias identificadas nos novos grupos homólogos distantes OrthoMCL/COMA e OMA/COMA (validados no SUPERFAMILY) com 3 organismos (core) é mostrada na Figura 4-16.

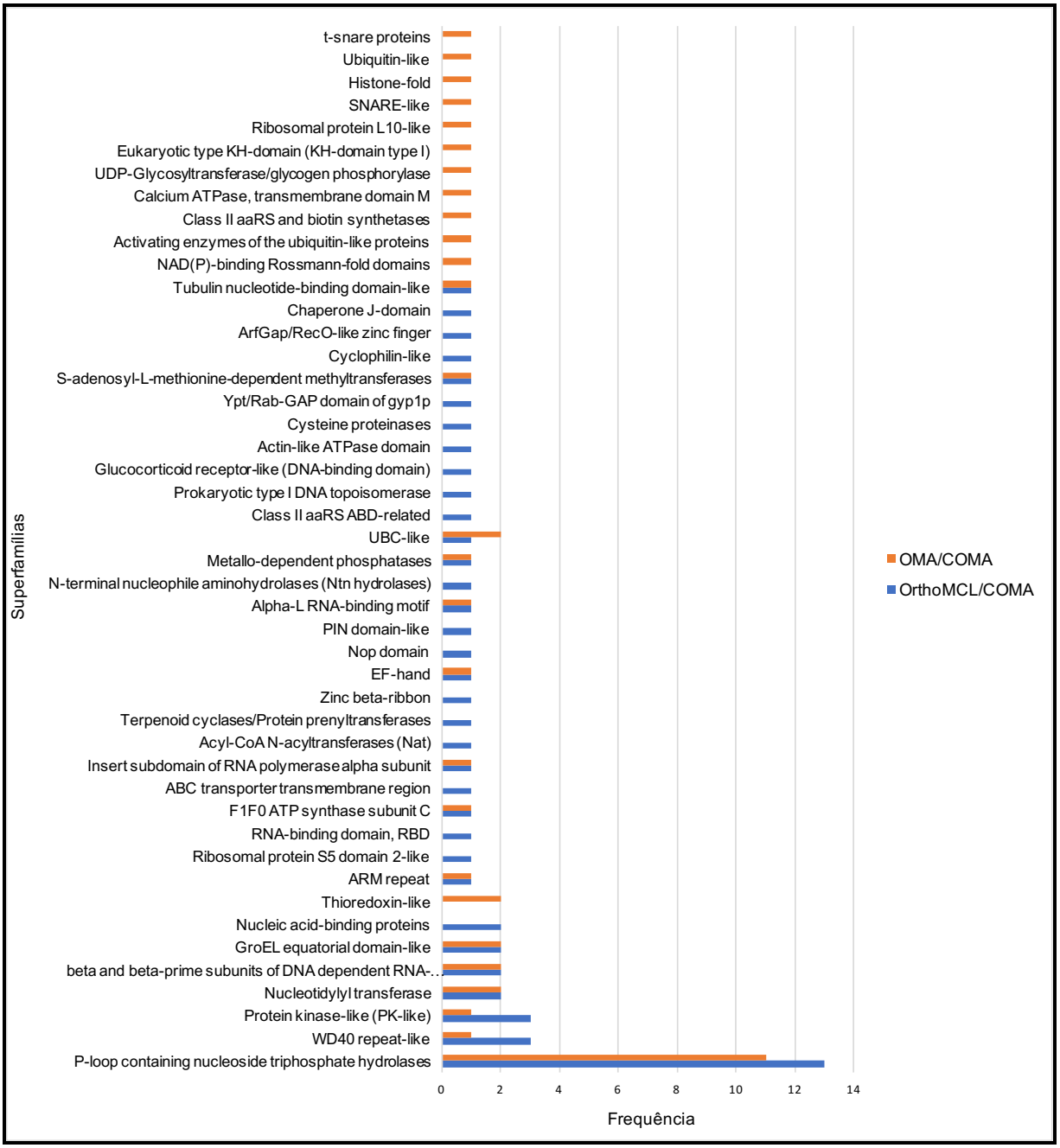



**Figura 4-16: Superfamílias identificadas nos novos grupos homólogos distantes OrthoMCL/COMA e OMA/COMA com ortólogos de *C. muris, E. invadens* e *T. grayi*.**

### 4.3.4 Inferência de distância evolutiva nos novos homólogos distantes

A análise dos resultados do programa Belvu mostrou que os novos homólogos distantes OrthoMCL/COMA e OMA/COMA apresentaram maiores distâncias evolutivas quando comparados ao conjunto total de grupos homólogos inferidos OrthoMCL e OMA (Figura 4-17). Em relação aos novos homólogos distantes OrthoMCL/COMA, a distância evolutiva média foi de 171,78 PAM, com uma distância evolutiva mínima de 27,15 PAM e distância evolutiva máxima de 300 PAM. Já em relação aos novos homólogos distantes OMA/COMA, a distância evolutiva média foi de 167,16 PAM, com uma distância evolutiva mínima de 77,83 PAM e distância evolutiva máxima de 300 PAM.

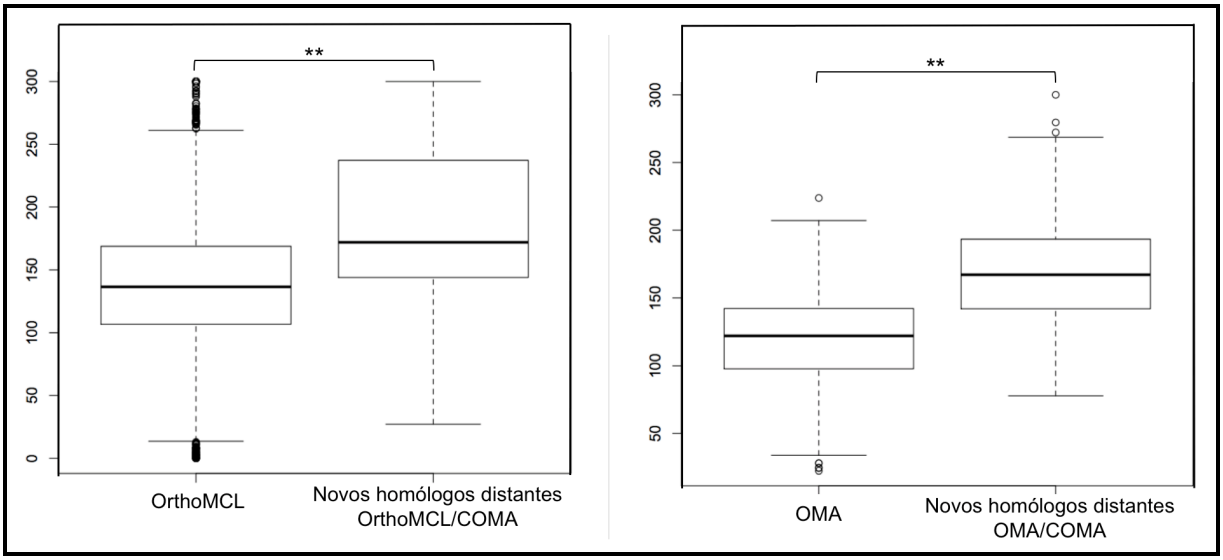



**Figura 4-17: Diagrama de caixa representando as distâncias evolutivas dos novos grupos homólogos distantes OrthoMCL/COMA e OMA/COMA e dos grupos OrthoMCL e OMA originais:**
As distâncias não apresentaram distribuição normal no teste de normalidade Shapiro-Wilk (OrthoMCL p-value: 2.2e-16; OMA p-value: 0.00000007573; Novos homólogos distantes OrthoMCL/COMA p-value: 0.08137; Novos homólogos distantes OMA/COMA p-value: 0.1237). A figura mostra que os Novos grupos homólogos distantes OrthoMCL/COMA apresentaram maiores distâncias estatisticamente significativas em relação aos do OrthoMCL: **p-value: 0.00000002897 e que os Novos grupos homólogos distantes OMA/COMA apresentaram maiores distâncias estatisticamente significativas em relação aos do OMA: **p-value: 3.509e-15 no resultado do teste Wilcoxon-Mann-Whitney. As linhas espessas dentro dos retângulos representam as respectivas medianas, os lados inferiores e superiores dos retângulos representam os primeiros e terceiros quartis, enquanto que as barras horizontais situadas nas extremidades representam os limites superiores e inferiores das respectivas distâncias.

# 5 DISCUSSÃO

## 5.1 Inferência de Grupos Homólogos em Protozoários Evolutivamente Distantes Com os Programas OrthoMCL e OMA

### 5.1.1 Inferência de homologia

Nossos resultados mostraram que o OMA apresentou um menor número de proteínas contribuindo para os grupos homólogos (10,72%) quando comparado ao OrthoMCL (56,06%) (Figura 4-1). É esperado um número maior de grupos homólogos no OrthoMCL em relação ao OMA, uma vez que o OMA infere apenas grupos ortólogos e o OrthoMCL também infere grupos parálogos. No entanto, mesmo removendo os grupos exclusivamente parálogos do OrthoMCL, este ainda inferiu mais grupos ortólogos: 1.481 (OrthoMCL) versus 1.231 (OMA). Tomando como referência apenas os grupos ortólogos compartilhados por todos os 3 organismos (core), o OrthoMCL também inferiu o maior número: 758, versus apenas 382 de OMA (Tabela 2).

Este resultado está em desacordo com um estudo anterior encontrado na literatura realizado por Dessimoz (Dessimoz et al., 2005), onde o OMA foi capaz de inferir um maior número de proteínas homólogas (66%) em comparação ao OrthoMCL (47%). Uma das causas para isto poderia ser o conjunto de dados usado por Dessimoz, com 150 espécies de grande biodiversidade, enquanto nesse experimento, usamos 3 protozoários de diferentes filos e, além disso, o algoritmo do OMA remove hits com alta distância evolutiva (Roth, Gonnet & Dessimoz, 2008).

Além disso, apesar de ter inferido um número menor de grupos homólogos, os resultados do OMA tiveram um alto nível de concordância com os resultados do OrthoMCL, existindo apenas 5,44% (67/1231) grupos homólogos exclusivos do OMA (Tabela 2).

Em relação às proteínas em grupos exclusivamente parálogos inferidos pelo OrthoMCL, nossos resultados mostraram que *E. invadens* tem a maioria deles: 57,1% (6.856/11.997), enquanto *T. grayi* tem 19,2% (2.034/10.583) e *C. muris* apenas 7.9% (309/3.934) (Tabela 2). Esta grande quantidade de proteínas parálogas em *E. invadens* está de acordo com a literatura, como já descrito por um estudo anterior (Ehrenkaufer et al., 2013), e corrobora com o fato de que *E. invadens* foi o

organismo que contribuiu com o maior número de proteínas em grupos de homólogos na inferência do OrthoMCL.

### 5.1.2 Inferência de distância evolutiva

Em nosso estudo, o OrthoMCL inferiu grupos ortólogos com maiores distâncias evolutivas do que o OMA. Isso provavelmente ocorre porque que o OMA remove resultados caracterizados por alta distância evolutiva (Roth, Gonnet & Dessimoz, 2008), o que torna este método desvantajoso quando o objetivo é inferir homólogos com grande distância evolutiva. Além disso, o OrthoMCL é um método probabilístico baseado em BLAST, que usa um clusterização de Markov para agrupar os melhores hits bidirecionais em grupos homólogos, método com potencial para inferir uma distância evolutiva mais alta (Trachana et al., 2011).

Vale salientar, no entanto, que grupos homólogos inferidos pelo OrthoMCL com ortólogos e parálogos, atingiram distâncias evolutivas de até 300 PAM. Isso pode ser explicado, provavelmente, pelo fato de que geralmente é esperado que parálogos sofram menos pressão evolutiva para manter a mesma função e e terminam divergindo mais do que ortólogos, (Forslund & Pekkari, 2011; Alberts et al., 2014).

### 5.1.3 Identificação de domínios conservados (CDD)

O uso de bases de dados de domínios conservados e famílias de proteínas tem se mostrado essencial para algumas análises de genômica comparativa que visam compreender o processo evolutivo, a interpretação dos dados genômicos e as relações filogenéticas entre as espécies. Nossos resultados mostram que apenas uma pequena parte das proteínas dos 3 organismos estudados tem mais de um domínio conservado reconhecido na base de dados CDD: 2,9% (113/3.934) em *C. muris*, 2,7% (319/11.997) em *E. invadens* e 2,4% (256/11.583) em *T. grayi* (Figura 4-3). Esta informação sobre a genômica funcional de espécies de protozoários está de acordo com relatórios anteriores na literatura. De acordo com Davidson (Davidson et al., 1003), os eucariotas geralmente têm uma grande quantidade de proteínas multidomínio. No entanto, dentre os eucariotos, os protozoários mostram o menor número dessas proteínas (Tordai et al., 2005). Além disso, o grande número de proteínas sem domínios conservados reconhecidos na base de dados CDD

(cerca de 42%) pode indicar que pode haver mais proteínas que são domínios múltiplos com domínios conservados ainda não identificados na base de dados CDD. Outra coisa que chama a atenção é que existem apenas 3 proteínas neste estudo com 5 domínios conservados identificados e 9 proteínas com 4 domínios conservados, todas os parálogas de *T. grayi*, não presentes em nenhum grupo OMA, mas pertencem ao mesmo grupo homólogo inferido por OrthoMCL. Nesse grupo homólogo OrthoMCL específico, existe também uma única proteína de *C. muris* com 2 domínios conservados identificados. Todas as proteínas desse grupo homólogo têm a mesma anotação: cadeia pesada de dineína (*dynein heavy chain*). A família da cadeia pesada de dineína é uma grande família de proteínas motoras; cujos estudos comparativos demonstraram que as isoformas equivalentes são bem conservadas entre as espécies. Algumas espécies apresentam formas únicas dessas proteínas (Asai & Wilkes, 2004), o que se reflete em nossas análises, onde foi encontrada uma quantidade diferente de domínios dentro da mesma família de proteínas, sugerindo que elas são homólogas pois têm domínios em comum, porém se diferenciam, ao terem diferentes quantidades de domínios conservados identificados.

### 5.1.4 Identificação de famílias de proteínas (Pfam-A)

As predições de família de proteínas da base de dados Pfam (Pfam-A) para os grupos homólogos inferidos nesta tese foram feitas utilizando o programa CLADE, ferramenta que mostrou em experimentos anteriores maior sensibilidade para a inferências desses domínios quando comparado com o programa HMMer, utilizado como padrão de buscas por inferência no portal da base Pfam (https://pfam.xfam.org) (Bernardes et al., 2015; 2016). Nossos resultados mostram que 20,2% (798/3.934) das proteínas de *C. muris*, 25,9% (3.117/11.997) das proteínas de *E. invadens* e 37,9% (4.010/10.583) das proteínas de *T. grayi* não possuem nenhuma família de proteínas identificada na base de dados Pfam-A (Figura 4-4). O número de proteínas que representam famílias distintas (com mais de um domínio Pfam-A) reconhecidas em nossa análise é o esperado para protozoários, de acordo com um estudo anterior (Tordai et al., 2005), onde os autores demonstraram que a proporção de proteínas com multidomínio Pfam-A diminui na ordem metazoa> plantas> fungos ~ protozoários> bactérias> arqueias e eles encontraram archaea mostrando apenas 23% das entradas com mais de um

domínio Pfam-A, enquanto que os metazoas apresentaram 39% das entradas correspondentes às proteínas com multidomínio Pfam-A. De acordo com eles, em Protozoários, a proporção de proteínas com multidomínio Pfam-A é de 32%, muito próximo dos nossos resultados, onde a soma das proteínas com 2, 3, 4 ou mais famílias de Pfam-A reconhecidas é de cerca de 29% (Figura 4-4). A família de proteínas mais comumente encontrada foi a Pkinase (PF00069), uma família de proteínas estruturalmente conservada, presente de *E. coli* à *H. sapiens*, e que desempenha um papel em uma infinidade de processos celulares, incluindo divisão, proliferação, apoptose e diferenciação (Manning et al., 2002). As proteínas sem hit no CDD, Pfam e outras bases de dados, ou seja, sem similaridade de sequência com proteínas conhecidas, podem ser potenciais órfãs (Fukuchi & Nishikawa, 2004), que são únicas para cada espécie e provavelmente desempenham um papel essencial no organismo de origem. Além disso, as proteínas encontradas em grupos homólogos compartilhados pelo menos por 2 espécies de protozoários podem ser, potencialmente, proteínas específicas de protozoários. Neste estudo, 2.621 proteínas homólogas não tiveram hit nas bases de dados CDD ou Pfam e podem ter, hipoteticamente, novos domínios conservados, essas proteínas representam 9,88% (2.621/26.514) do conjunto de dados usado neste estudo e a 17,24% (2.621/15.202) de todos as proteínas homólogas inferidas e somente 1,9% (50/2.621) das proteínas com novos domínios conservados hipotéticos pertencem a grupos homólogos compartilhados por pelo menos 2 organismos.

## 5.2 Inferência de homologia baseada em uma abordagem de reconciliação para a genômica comparativa de protozoários

### 5.2.1 Algoritmo de Reconciliação – Inferência de Supergrupos

Os Supergrupos homólogos gerados a partir da nossa abordagem de reconciliação (Figura 3-4) não poderiam ter sido inferidos por nenhuma das 2 ferramentas de inferência de homologia devido às diferentes abordagens usadas pelos dois programas: seus pontos de corte e os algoritmos utilizados para inferir homólogos. Em seus primeiros passos, o algoritmo OMA usa um score > 85 para excluir alinhamentos que não considera significativos e portanto, pequenas sequências ficam fora de sua análise (Roth, Gonnet & Dessimoz, 2008), e além disso, remove do gráfico inicial os melhores Hits bidirecionais com alta distância evolutiva (Trachana et al., 2011). Por outro lado, o algoritmo OrthoMCL usa e-value 1E-05 como parâmetro de corte padrão e reconhece grupos homólogos por pontuação (Best Hits) baseados em BLAST, que usa clusterização de Markov como parte do seu pipeline. (Margelevicius & Venclovas, 2010). Essas diferenças fundamentais entre os dois programas podem explicar por que algumas proteínas não estão incluídas nos grupos homólogos inferidos por cada um dos dois programas, e o por que foi possível inferir os Supergrupos neste estudo, reconciliando os grupos homólogos inferidos pelo OMA e OrthoMCL.

Os fragmentos de sequência conservados de forma evolutiva (ou domínios) são bons indicadores de homologia, pois podem fornecer caracterização funcional com base na presença de padrões de assinatura de sequência e podem servir como ponto de partida para a anotação e classificação funcional (Marchler-Bauer et al., 2015) e o mesmo pode ser dito sobre as famílias e superfamílias de proteínas, uma vez que são conjuntos de sequências evolutivamente relacionadas (Finn et al., 2016).

O alinhamento múltiplo de sequências (MSA) é uma ferramenta de fundamental importância para a biologia evolutiva e molecular que visa a identificação de domínios funcionais, predição de estruturas e inferência de origem evolutiva comum, revelando histórias evolutivas de famílias protéicas (Pais et al., 2014). Uma vez que a análise da conservação de sequências tem sido utilizada como ferramenta para a inferência de homologia nas espécies de *Entamoeba histolytica* (Cuadrat et al., 2014), *Trypanosoma cruzi* (Jackson et ai., 2016) e

*Cryptosporidium hominis* (Xu et al., 2004), produzimos alinhamentos múltiplos para todos os Supergrupos inferidos com o objetivo de entender melhor seu nível de conservação. Os alinhamentos múltiplos desse estudo foram produzidos utilizando a ferramenta MAFFT, visto que esse método foi bem recomendado para melhor precisão de alinhamentos de múltiplas sequências em trabalhos anteriores (Pais et al., 2014). O tamanho dos alinhamentos múltiplos dos Supergrupos variou de 150 a 1,311 aminoácidos apresentando mais de 34,51% da identidade média e 27,71% da identidade média entre as sequências de maior distância de cada Supergrupo. É esperado que homólogos distantes tenham identidade média baixa, sendo uma das causas de que a simples comparação de sequências não seja considerada uma boa estratégia para inferir homólogos com grande distância evolutiva (Margelevičius & Venclovas, 2010).

Considerando a validação CDD e Pfam-A dos Supergrupos, o baixo nível de identidade entre as sequências de cada Supergrupo sugere que eles podem ser considerados homólogos distantes, que não poderiam ser inferidos por OrthoMCL ou OMA, o que também sugere a análise das distâncias evolutivas dos grupos OMA, OrthoMCL e Supergrupos que mostrou que, entre estes, os Supergrupos são os homólogos mais distantes (Figura 4-9) e que a diferença entre as distâncias é estatisticamente significativa via teste Wilcoxon-Mann-Whitney (OrthoMCL / Supergroups: p-value 0.0005238 ; OMA / Supergroups: p-value: 0.0000004886) (Figura 4-9).

A filogenia tem sido usada como uma ferramenta para inferência definitiva de homologia, inclusive nas espécies *Entamoeba histolytica* (Cuadrat et al,. 2014), *Trypanosoma cruzi* (Jackson et al., 2016) e *Cryptosporidium hominis* (Xu, Widmer & Wang, 2004). Nós também usamos filogenia em nosso estudo para nos ajudar a validar os nossos Supergrupos homólogos. Alinhamentos múltiplos puderam ser inferidos para todos os Supergrupos validados por domínio conservado ou família de proteínas (59/59), porém, para 18,65% (11/59) dos Supergrupos não foi possível inferir árvores filogenéticas baseadas em Máxima verossimilhança usando o RAxML (Stamatakis et al., 2014), uma vez que esses 11 Supergrupos têm apenas 3 proteínas cada e essa ferramenta, devido a restrições em seu algoritmo, não é capaz de inferir árvores filogenéticas com menos de 4 sequências. Nossos resultados também mostraram que mais de 80% dos nossos Supergrupos

homólogos são equivalentes às entradas SUPERFAMILY e mais de 77% às entradas Pfam Clans. Nesse estudo, esses indicadores de homologia foram utilizados para validar a inferência dos Supergrupos homólogos. Baseada nestes resultados, a Tabela 6 enumera alguns métodos de inferência de homologia distante.

**Tabela 6: Comparação entre alguns métodos de inferência de homologia distante.**

| SUPERFAMILY | Pfam clans | Supergrupos |
|---|---|---|
| Baseado em uma coleção de modelos ocultos de Markov (HMM), que representam domínios de proteínas estruturais no nível de superfamília do SCOP | Baseado na presença de estruturas relacionadas e resultados significativos da comparação pHMM-pHMM | Baseado na reconciliação dos resultados de outros programas; usando como critério de validação a presença do (a) mesmo domínio conservado (CDD) ou (b) mesma família de proteínas (Pfam-A) em todas as proteínas de um grupo. |

### 5.2.2 Análises funcionais

Análises posteriores foram realizadas visando aprofundar o conhecimento sobre a importância e funções biológicas dos Supergrupos homólogos inferidos, como segue:

(i) Categorização funcional, que é uma ferramenta muito útil para a genômica comparativa, uma vez que tem sido amplamente utilizada em muitas espécies, como *L. amazonensis* (Tschoeke et al., 2014), *E. histolytica*, *P. falciparum*, *L. major*, *T. brucei* e *T. cruzi* (Cuadrat et al,. 2014). Nas proteínas dos Supergrupos inferidos em nosso estudo, a categoria funcional mais comum encontrada é a "S" (Função desconhecida), esta categoria funcional é usada para famílias de proteínas que incluem pelo menos 100 proteínas de pelo menos dois filos diferentes (Galperin et al., 2014), o que é indicação de que eles são homólogos com função ainda não inferida. Além disso, vale destacar que os homólogos de proteínas para os quais as funções biológicas permanecem desconhecidas é vital para o progresso da anotação do genoma (Galperin & Koonin, 2010). Esses grupos com função desconhecida podem precisar de mais estudos para definir suas funções, apesar de terem formado grupos homólogos e serem preservados, o que pode indicar que eles

têm uma função relevante para esses organismos. Mesmo com as evidências geradas pela análise de família de proteínas, de domínios conservados e filogenia, o fato de que as proteínas de 2 Supergrupos tenham sido identificadas como pertencentes a 2 categorias funcionais distintas (SG_1247 com categorias "P" e "U" e SG_1633 com categorias "J" e "O") também pode sugerir que esses 2 supergrupos sejam homólogos distantes.

(ii) A inferência da via metabólica do KEGG, que pode ser utilizada como referência para a reconstrução funcional em grupos homólogos (Ogata et al., 1998). Nesse estudo, 6 Supergrupos foram escolhidos como estudo de caso, especificamente por fazerem parte da via “metabolismo” do KEGG nos organismos estudados (Tabela 4):

(a) o SG_1364 participa da via de fosforilação oxidativa (Oxidative phosphorylation pathway) com a enzima pirofosfatase inorgânica (EC 3.6.1.1) que catalisa a conversão de uma molécula de pirofosfato em dois fosfatos. Esta enzima, amplamente distribuída entre bactérias, fungos, protozoários e algas (Motta et al., 2004), desempenha um papel essencial no metabolismo lipídico (Ko et al., 2007);

(b) o SG_1363 participa do metabolismo da histidina (Histidine metabolism) com a enzima carnosina N-metiltransferase (EC 2.1.1.22). A identificação do gene que codifica a carnosina N-metiltransferase pode ser benéfica para a inferência das funções biológicas da anserina (Drozak et al., 2013). A Anserina (β-alanil-N-π-metil-L-histidina) é um derivado natural da carnosina (β-alanil-L-histidina) que foi relatado estar presente no sistema nervoso central e no músculo esquelético de muitos vertebrados (Quinn et al,. 1992), apesar disso, sua função fisiológica permanece desconhecida (Drozak 2015 et al., 2015);

(c) o SG_1634 participa das vias do sistema de retransmissão de enxofre (Sulfur relay system pathway) e do metabolismo de tiamina (Thiamine metabolism pathway) com a enzima cisteína dessulfurase (EC: 2.8.1.7) que catalisa a reação química L-cisteína + [enzima] -cisteína <=> L-alanina + [enzima] -S-sulfanilcisteína. Em *T. brucei*, esta enzima está envolvida na biossíntese de aglomerados de ferro-enxofre, tio-nucleósidos em tRNA, biotina, lipoato, tiamina e piranopterina (molibdopterina) (Poliak et al,. 2010);

(d) o SG_843 participa das vias do metabolismo de ácido graxo (Fatty acid metabolism pathway), do alongamento do ácido graxo (Fatty acid elongation

pathway), da biossíntese do hormônio esteróide (Steroid hormone biosynthesis pathway) e da biossíntese das vias de ácidos graxos não saturados (Biosynthesis of unsaturated fatty acids pathway) com (1) enzima 17-beta-estradiol 17-desidrogenase (CE: 1.1.1.62) que participa do colesterol pós-qualitano Biossíntese em mamíferos (Marijanovic et al., 2003) e (2) enzima 3-oxoacil-CoA redutase de cadeia muito longa (EC: 1.1.1.330), que é um constituinte essencial de células eucarióticas, mais comumente encontradas como blocos de construção de esfingolípidos, além disso, eles são também componentes importantes de glicerofosfolípidos, ésteres de esterol, triacilgliceróis e ésteres de cera (Beaudoin et al., 2009);

(e) SG_1241, participa das vias de biossíntese de aminoácidos (Biosynthesis of amino acids pathway), metabolismo de glicina, serina e treonina (Glycine, serine and threonine metabolism pathway), metabolismo de cisteína e metionina (Cysteine and methionine metabolism pathway), metabolismo de carbono (Carbon metabolism pathway), e de biossíntese de aminoácidos (Biosynthesis of amino acids pathway), metabolismo de enxofre (Sulfur metabolism pathway). Note-se que este Supergrupo (SG_1241), mesmo tendo hit com mais de um KO, e consequentemente com enzimas distintas: enzima de cistatina-beta-sintase (EC: 4.2.1.22) e Cisteína sintase (EC: 2.5.1.47), respectivamente, essas 2 enzimas pertencem à mesma via: síntese de precursor de tripanotiona (trypanothione precursor synthesis) em espécies de Tripanosomatídeos (Beltrame-Botelho et al., 2016) e ambas podem catalisar reações semelhantes (adicionando sulfato de hidrogênio a L-Serina ou O-Acetil-L-serina) (KEGG REACTION: R00897); ( KEGG REACTION: R00891);

(f) O SG_711 participa das vias do metabolismo da cisteína e da metionina, do metabolismo de selenocompostos (Selenocompound metabolism pathway) e da Biossíntese de aminoácidos (Biosynthesis of amino acids pathway), com a enzima beta-liasa de cistatina (EC: 4.4.1.8), encontrada em plantas, bactérias e fungos, é uma Parte essencial da via de biossíntese de metionina e a ausência desta enzima em organismos superiores torna o alvo importante para o desenvolvimento de antibióticos e herbicidas (Breitinger et al.,2001).

## 5.3 Inferência e análise de homólogos distantes em Protozoa por meio da comparação pHMM - pHMM (perfis de Modelo Oculto de Markov) visando a identificação de superfamílias

### 5.3.1 Comparação pHMM - pHMM com o COMA

Quando uma busca de similaridade de sequências encontra uma correspondência estatisticamente significativa, podemos inferir com confiança que estas sequências são homólogas; mas não podemos ter certeza de que nenhum homólogo está presente se nenhuma correspondência estatisticamente significativa for encontrada em uma base de dados (Pearson, 2013). Vê-se necessário a utilização de métodos que identifiquem homólogos distantes, que possuem baixa similaridade de sequências, porém com semelhança estrutural estatisticamente significativa ou forte similaridade de sequência com uma sequência intermediária (Pearson, 2013). Genes homólogos distantes são potencialmente essenciais, justificando o fato de se manterem conservados mesmo em organismos evolutivamente distantes.

Visto que a comparação entre perfis do modelo oculto de Markov (pHMM - pHMM) é reconhecida como uns dos métodos de inferência de homologia mais sensíveis da atualidade (Remmert et al., 2011), este trabalho teve como objetivo identificar homólogos distantes e verificar a quais superfamílias de proteínas estes poderiam estar associados. Os grupos homólogos inferidos pelo OrthoMCL e pelo OMA serviram como entrada para inferir os melhores hits recíprocos (pHMM – pHMM) utilizando o programa COMA, e assim verificar a existência de novos grupos homólogos distantes, que não poderiam ter sido identificados pelos algoritmos do OrthoMCL e do OMA. Além disso, foi nosso objetivo identificar as funções biológicas desses homólogos distantes nos organismos estudados.

Segundo Margelevičius & Venclovas (2010), o algoritmo do COMA além de realizar a comparação pHMM - pHMM, calcula penalidades para gap dependendo de posição e usa um sistema de score global, que segundo o autor, permite uma solução analítica dos parâmetros estatísticos na estimativa de significância de um hit. O autor demonstra resultados mais sensíveis em relação com outros métodos estado-da-arte como COMPASS, HHsearch e o PSI- BLAST (Margelevičius & Venclovas, 2010).

Nota-se que os resultados do programa OMA, mesmo com menos grupos ortólogos inferidos em relação ao do OrthoMCL (OrthoMCL: 1.481, OMA: 1.231) (Tabela 2), serviram como base para uma quantidade maior de hits recíprocos pHMM – pHMM identificados no geral, quando comparados aos gerados a partir dos resultados do OrthoMCL (OrthoMCL: 74, OMA: 128) porém tiveram menos hits recíprocos pHMM – pHMM identificados entre os grupos ortólogos com 3 organismos (core) (OrthoMCL: 57, OMA: 46) (Figura 4-12 e Figura 4-13), e isso pode se dever ao algoritmo do OMA, que retira homólogos distantes em um dos seus passos iniciais (Roth, Gonnet & Dessimoz, 2008) e ao uso dométodo de comparação pHMM – pHMM, que tem maior potencial de identificação de homólogos distantes (Remmert et al., 2011) pode ter identificado homólogos mais distantes, inicialmente descartados pelo OMA. Além disso, o fato de OMA não ter inferido parálogos pode ter influenciado nesse resultado, visto que proteínas parálogas mais distantes evolutivamente podem ter sido agrupadas em grupos ortólogos distintos por esse programa. A título de comparação, 32,1% (987/3.092) dos grupos homólogos inferidos pelo OrthoMCL, que foram utilizados como entrada para a comparação pHMM – pHMM, são grupos homólogos com proteínas ortólogas e parálogas.

### 5.3.2 Validação dos homólogos distantes inferidos pelo COMA

Os grupos OMA e OrthoMCL que tiveram melhor hit recíproco pHMM – pHMM no COMA, foram analisados no portal SUPERFAMILY (Supfam.org) e, confirmando a nossa hipótese de que eles são homólogos distantes, a maior parte deles (Figura 4-11), (Figura 4-12) e (Figura 4-13) demonstraram estar associados à mesma superfamília, sendo validados e aqui chamados de Novos grupos homólogos distantes OrthoMCL/COMA e OMA/COMA.

Nota-se que a maioria dos melhores hits recíprocos validados ocorreu entre os ortólogos com os 3 organismos (core) tanto para os grupos OMA com 91,3%(42/46) (Figura 4-13) quanto para os grupos OrthoMCL com 94,73%(54/57) (Figura 4-12), uma provável causa disso seria que os homólogos entre essas 3 espécies pertencentes a filos diferentes e portanto, de distintas regiões da árvore da vida (Figura 3-1), tendem a ser as regiões mais conservadas entre os seus genomas (Koonin, 2005).

Por outro lado, os grupos distantes prováveis com melhores hits recíprocos pHMM – pHMM que menos puderam ser validados via identificação de superfamílias são os que utilizaram os parálogos inferidos pelo OrthoMCL (Figura 4-11) como entrada, esse resultado pode ser parcialmente explicado porque grupos parálogos tendem a divergir mais do que grupos ortólogos (Alberts et al., 2014, Forslund & Pekkari, 2011).

Os novos grupos homólogos distantes OrthoMCL/COMA e OMA/COMA apresentaram maior distância evolutiva em relação a todos os grupos OrthoMCL e OMA originais, e análises estatísticas posteriores (teste Wilcoxon-Mann-Whitney) mostraram que a diferença entre estas distâncias é estatisticamente significativa (OrthoMCL / Novos homólogos distantes OrthoMCL/COMA: p-value 0.00000002897 ; OMA / Novos homólogos distantes OMA/COMA: p-value: 3.509e-15) (Figura 4-17), essa informação também sugere que os novos homólogos distantes inferidos via concatenação dos grupos OrthoMCL e dos grupos OMA com melhor hit recíproco pHMM - pHMM são homólogos distantes.

Sobre os homólogos distantes prováveis (melhores hits recíprocos pHMM – pHMM (COMA) que não foram validados pelo SUPERFAMILY):

Para analisar as relações entre os homólogos distantes prováveis com 3 organismos (core), foram criados 2 grafos, demonstrando, quando possível, até 3 melhores hits desses grupos (para aqueles que tiveram mais de 1 hit), (Figura 4-14, Figura 4-15). Nota-se ligações entre um número maior de grupos com a mesma superfamília identificada, formando conjuntos maiores de grupos homólogos distantes, o que é indício de que esses grupos são verdadeiramente homólogos distantes. Além disso, as relações identificadas em alguns grupos com melhor hit recíproco pHMM - pHMM e sem superfamília identificada, tanto nos resultados do OrthoMCL quanto do OMA, indicam que estes são grupos ortólogos com domínios ainda não anotados no SUPERFAMILY (Figura 4-14, Figura 4-15).

Entre os homólogos distantes prováveis OrthoMCL/COMA com 3 organismos, apenas 1,75% (1/57) não puderam ser validados por terem melhores hits recíprocos associados a superfamílias distintas na base de dados SUPERFAMILY, um resultado próximo ao obtido entre os homólogos distantes prováveis OMA/COMA com 3 organismos, onde 4,34% (2/46) não puderam ser validados pelo mesmo motivo. Neste caso, existe a probabilidade de algumas dessas proteínas serem

multidomínios, com domínios conservados ainda não anotados na base de dados SUPERFAMILY, pois esses grupos foram inferidos por programas para inferência de homologia (OrthoMCL e OMA) e também tiveram melhor hit recíproco pHMM – pHMM (COMA), sendo estes indícios de que realmente se tratam de grupos homólogos.

### 5.3.3 Análises funcionais das superfamílias a serem associadas aos grupos homólogos OrthoMCL/OMA alvos de reanotação funcional

Nossos resultados mostraram que 6 prováveis homólogos distantes não puderam ser validados porque tiveram superfamília identificada em apenas 1 dos 2 grupos homólogos com melhor hit recíproco pHMM – pHMM (ou seja, apenas 1 grupo teve superfamília identificada, como apresentado nas figuras Figura 4-11, Figura 4-12 eFigura 4-13), indicando a possibilidade de transferência de anotação nestes casos (Tabela 5). Análises funcionais das superfamílias que podem ser associadas a grupos homólogos neste estudo são mostradas abaixo:

(a) A superfamília “Class II aaRS ABD-related” é formada por 2 famílias de proteínas e é uma enzima chave durante o processo de biossíntese protéica, contendo um domínio central catalítico, responsável pela ativação do aminoácido e um domínio de ligação do anticódon ao seu tRNA cognato. Quando essa proteína é de Classe II, possui as sintetases específicas para alanina, asparagina, ácido aspártico, glicina, histidina, lisina, fenilalanina, prolina, serina e treonina (Tang & Huang, 2005). Por se tratar de uma proteína multi-domínio muito antiga, a superfamília “Class II aaRS ABD-related” pode ser encontrada entre os três domínios da vida (Smith & Hartman, 2015).

(b) A superfamília “ARM repeat” ou “armadillo repeat” é formada por 13 famílias de proteínas e é um motivo característico de aproximadamente 40 aminoácidos repetidos compartilhado por proteínas com diversos papéis celulares, incluindo sinalização intracelular e regulação do citoesqueleto (Hatzfeld, 1999). Um subconjunto dessas proteínas é conservado entre os reinos eucarióticos (Coates, 2003). Cada repetição deste motivo é composta por um par de hélices alfa que formam uma estrutura em gancho e a conformação tridimensional de uma “ARM repeat” é conhecida a partir da estrutura cristalina da beta-catenina, onde as 12

repetições formam uma super-hélice de hélices alfa com três hélices por unidade (Breitinger, Nelson & Weis, 2001).

(c) A superfamília “PH domain-like” ou “Pleckstrin homology domain-like” é formada por 13 famílias de proteínas e possui pequenos domínios modulares que ocorrem em uma grande variedade de proteínas de sinalização, onde servem como domínios de ligação para lipídios (Cozier et al., 2004). Este domínio está presente, por exemplo, na proteína “Pleckstrin”, encontrada nas plaquetas sanguíneas do humano e do rato (NIH, 1983), o que corrobora com nossos resultados. Esses domínios têm uma topologia de barril beta parcialmente aberta que é limitada por uma hélice alfa. A estrutura dos domínios de PH é semelhante ao domínio de ligação à fosfotirosina (PTB) encontrado no IRS-1 (substrato 1 do receptor de insulina) (Eck et al., 1996).

(d) A superfamília “GTPase activation domain, GAP” é formada por 2 famílias de proteínas e é uma Superfamília de proteínas reguladoras cujos membros podem se ligar a proteínas G ativadas e estimular sua atividade GTPase, com o resultado de terminar o evento de sinalização. Como as proteínas G estão envolvidas em vários processos celulares importantes, sua regulação tem importância fundamental (Krauss, 2014).

(e) A superfamília “N-terminal nucleophile aminohydrolases”, também conhecida como “hidrolases Ntn” é formada por diversas enzimas e possui 7 famílias de proteínas, sendo formado por dois grupos catalíticos: nucleófilo e doador de prótons. As hidrolases Ntn utilizam a cadeia lateral do resíduo amino-terminal, incorporado em uma folha beta, como o nucleófilo no ataque catalítico no carbono carbonílico. O nucleófilo é a cisteína no GAT, a serina na penicilina acilase e a treonina no proteassoma. Todas as enzimas compartilham uma conformação não convencional na qual o nucleófilo e outros grupos catalíticos ocupam sítios equivalentes. Esta conformação fornece tanto a capacidade de ataque nucleofílico como a possibilidade de processamento autocatalítico (Brannigan et al., 1995). Estas enzimas catalisam a hidrólise da ligação amida (Oinonen & Rouvinen, 2000) e encontram-se amplamente distribuídas entre os organismos procarióticos e eucarióticos (Bompard-Gilles et al., 2000).

(f) A análise funcional da superfamília “P-loop containing nucleoside triphosphate hydrolases” é mostrada no tópico 5.3.4 desta tese.

### 5.3.4 Análises funcionais das superfamílias mais frequentemente identificadas nos novos grupos homólogos distantes (core)

Visando identificar as funções biológicas dos novos homólogos distantes inferidos via concatenação dos melhores hits recíprocos pHMM – pHMM entre os grupos OMA e OrthoMCL com 3 organismos (core), foram realizadas análises funcionais das superfamílias mais frequentemente identificadas na validação dos mesmos (Figura 4-16).

(a) A superfamília encontrada com maior frequência em nossas análises, tanto nos melhores hits do OrthoMCL quanto nos do OMA é a superfamília “P-loop containing nucleoside triphosphate hydrolases” (P-loop NTPase), que foi identificada e utilizada como parâmetro de validação de 24 melhores hits recíprocos pHMM – pHMM (13 do OrthoMCL e 11 do OMA). Esta superfamília é formada por 24 famílias de proteínas e, segundo sua anotação funcional, trata-se de uma “pequena molécula de ligação” que está presente em diversos processos biológicos, tais como: resposta a estímulos, organização de componentes celulares ou biogênese, resposta a ferimentos, etc. (Gough et al., 2001). A reação mais comum catalisada por enzimas do enovelamento da P-loop NTPase é a hidrólise da ligação de fosfato de beta-gama de uma ligação de nucleosídeo trifosfato (NTP) (Leipe, Koonin & Aravind, 2004). A energia da hidrólise de NTP é tipicamente utilizada para induzir alterações conformacionais em outras moléculas, o que constitui a base das funções biológicas da maioria das P-loop NTPase. As P-loop NTPase mostram preferência substancial pelo substrato por ATP ou GTP (Leipe, Koonin & Aravind, 2004). O fato de a superfamília P-loop NTPase ter sido a mais identificada em nossas análises, corrobora com o estudo prévio feito por Leipe (Leipe, Koonin & Aravind, 2004), onde o autor afirma que o enovelamento da P-loop NTPase é o domínio mais prevalente entre os vários e distintos enovelamentos existentes em proteínas de ligação codificadas nos genomas da maioria das formas de vida celular.

(b) Outra superfamília que aparece entre as mais frequentemente identificadas é a “Nucleotidylyl transferase”, utilizada como parâmetro de validação para 4 melhores hits recíprocos pHMM – pHMM (2 de cada ferramenta). Esta

superfamília é formada por 5 famílias de proteínas e, segundo sua anotação funcional, é responsável por transporte e metabolismo de nucleotídeos e também está presente em diversos processos biológicos, tais como: expressão gênica, processo de metabolismo de tRNA, atividade de ligases, etc. (Gough et al., 2001). Estas enzimas foram mais estudadas em plantas (Khan et al., 2015), bactérias (Xu et al., 2010) e fungos (Ullrich et al., 2001) mas possuem homólogos maioria das formas de vida celular (Gough et al., 2001).

(c) A superfamília "beta and beta-prime subunits of DNA dependent RNA-polymerase" também foi identificada e utilizada como parâmetro de validação para validar 4 melhores hits recíprocos pHMM – pHMM (2 de cada ferramenta). Esta superfamília, cujo o site catalítico é formado pela associação de dois domínios de barril beta de duplo psi, um de cada subunidade (Murzin et al., 2009), é formada por 2 famílias de proteínas e, segundo sua anotação funcional, ela está relacionada à expressão gênica, particularmente na etapa de transcrição, estando conservada entre os reinos eucarióticos e procarióticos (Gough et al., 2001).

(d) Outra superfamília que foi identificada e utilizada como parâmetro de validação para 4 melhores hits recíprocos pHMM – pHMM (3 do OrthoMCL e 1 do OMA) é a "WD40 repeat-like" (repetição de WD ou beta-transducina), formada por 2 famílias de proteínas. Segundo sua anotação, ela é um motivo estrutural curto, altamente conservado, com aproximadamente 40 aminoácidos geralmente terminado em um dipéptido de ácido triptofano-ácido aspártico (W-D) e encontrado em todos os eucariotos, mas não em procariotos (Neer et al., 1994). Existe um crescente interesse em proteínas WD40 repeat-like devido à sua associação com histonas e nucleossomas e por suas funções em uma variedade de complexos de modificação de histonas/cromatina (Suganuma, Pattenden & Workman, 2008). Esta superfamília está envolvida no processo biológico de organização ou biossíntese celular, alquilação de proteínas, modificação de histona, etc.

(e) A superfamília "GroEL-like equatorial domain" foi identificada e utilizada como parâmetro de validação para 4 melhores hits recíprocos pHMM – pHMM (2 de cada ferramenta), sendo constituída de 2 famílias protéicas: grupo I, encontrado em eubactérias (por exemplo, GroEL em Escherichia coli) e organelas eucarióticas de descendência eubacteriana (por exemplo, Cpn60 em mitocôndrias e cloroplastos) e grupo II, encontrado em archaea e no citosol eucariótico (complexo CCT ou TCP-1) (Kusmierczyk et al., 2001; Kubota et al., 1995 ). E possui dois subdomínios de 4

helicoidais que estão relacionados por uma pseudo-díade que passa pela região de ligação ATP, estando assim envolvida no processo biológico de duplicação gênica (Murzin et al., 2009).

(f) A superfamília “Protein kinase-like (PK-like)” também identificada e utilizada como parâmetro de validação para 4 melhores hits recíprocos pHMM – pHMM (3 do OrthoMCL e 1 do OMA). Esta superfamília é formada por 7 famílias e compartilha semelhanças funcionais e estruturais com a conformação de ATP e PIPK (Murzin et al., 2009) e está envolvida no processo biológico de regulação gênica, catalisando o processo de fosforilação, um dos principais participantes nos mecanismos de regulação das proteínas. (Engh et al., 2002). Esta enzima é comum no genoma humano, que possui cerca de 560 genes de proteínas kinase, constituindo cerca de 2% de todos os genes humanos (Manning et al., 2002). As proteínas kinase eucarióticas são enzimas que pertencem a uma família muito extensa de proteínas que compartilham um núcleo catalítico conservado (Hanks, 2003).

Por essas análises funcionais, nota-se que a maioria das superfamílias de proteínas identificadas nos novos grupos homólogos distantes são relacionadas a funções celulares fundamentais.

# 6 CONCLUSÕES

A comparação entre o OrthoMCL (grupos exclusivamente ortólogos) e o OMA mostrou que o primeiro é capaz de inferir grupos ortólogos com maior distância evolutiva do que o segundo. Por outro lado, a maior estringência do OMA ao excluir homólogos mais distantes, torna seus resultados mais confiáveis.

A maioria das proteínas dos genomas dos 3 protozoários estudados não teve mais de 1 Domínio Conservado (CDD) (Figura 4-3) ou Família de proteínas (Pfam-A) (Figura 4-4) reconhecido ou caracterizado nas bases de dados especializadas. Além disso, em nossas análises, grande parte desses genomas não teve nenhum CDD ou Pfam-A identificado.

O nosso método de reconciliação foi capaz de conciliar a inferência de homologia de outras 2 metodologias distintas, agrupando grupos homólogos OMA e OrthoMCL em Supergrupos homólogos.

Estes Supergrupos homólogos não poderiam ser inferidos separadamente pelo OMA ou pelo OrthoMCL.

Os resultados apresentados provavelmente são subestimados, uma vez que (a) nos organismos estudados, ainda existe uma ampla gama de proteínas sem família de proteínas (Figura 4-4) ou domínios conservados (Figura 4-3) identificados e (b) proteínas com diferentes famílias de proteínas ou diferentes domínios conservados podem ser proteínas com multi-domínios ainda não identificados.

Comparações com as bases de dados SUPERFAMILY e Pfam clans indicam que os Supergrupos homólogos são homólogos distantes (Tópico 41).

Os Supergrupos homólogos apresentaram maior distância evolutiva quando comparados com os grupos OrthoMCL e OMA originais (Figura 4-9).

Sendo assim, os Supergrupos homólogos inferidos nesta tese podem ser comparados com os grupos Superfamily e Pfam clans, os três representando grupos homólogos distantes (Tabela 6).

Até onde sabemos, este é o primeiro estudo a propor essa abordagem de reconciliação, tornando possível a inferência dos Supergrupos homólogos.

Sendo uma das principais contribuições desta tese, nosso método, pode ser utilizado como suporte para o estudo de qualquer espécie, com outros programas para inferência de homologia usados como entrada no processo.

Este estudo foi publicado na revista *Evolutionary Bioinformatics* sob o título: “Homology Inference Based on a Reconciliation Approach for the Comparative Genomics of Protozoa”.

A comparação entre perfis do modelo oculto de Markov (pHMM – pHMM) foi capaz de inferir grupos homólogos distantes, não identificáveis separadamente nos resultados no OrthoMCL ou do OMA, sendo a grande maioria destes validados pela identificação de superfamílias na base de dados SUPERFAMILY. Sua maior distância evolutiva pôde ser comprovada neste experimento (Figura 4-17).

Os grupos homólogos inferidos pelo programa OMA serviram como base para a obtenção de uma maior quantidade de grupos homólogos distantes quando comparados aos grupos homólogos inferidos pelo OrthoMCL, pois os pHMM gerados a partir desses grupos tiveram uma maior frequência de melhores hits recíprocos. Por outro lado, os grupos homólogos inferidos pelo programa OrthoMCL serviram como base para uma maior quantidade de grupos homólogos distantes entre os 3 organismos estudados (core).

A estratégia do uso do COMA para comparação pHMM – pHMM entre os resultados do OrthoMCL e do OMA, demonstrou a possibilidade de reanotação de grupos homólogos sem superfamília associada na base de dados SUPERFAMILY (Tabela 5).

A maioria das superfamílias de proteínas identificadas são relacionadas a funções celulares fundamentais (Figura 4-16).

# 7 PERSPECTIVAS

Esses resultados encorajadores servem de base para a automação futura do algoritmo de reconciliação de Supergrupos homólogos, provavelmente na forma de um pipeline ou workflow.

# 8 REFERÊNCIAS BIBLIOGRÁFICAS


1. Ruppert EE, Fox RS, Barnes RD. Invertebrate Zoology: A Functional Evolutionary Approach. Syst Biol. 2004;

2. Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, et al. Molecular Biology of the Cell 6e. Vol. 6, Garland Science. 2014. 1465 p.

3. Sleigh MA. The biology of Protozoa. Cambridge University Press. Chichester, UK: John Wiley & Sons, Ltd; 1973.

4. Gransden WR. Topley and Wilson's Microbiology and Microbial Infections, 9th edition (CD-ROM) [Internet]. Vol. 52, Journal of Clinical Pathology. 1999. 237-238 p. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC501099/%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC501099/pdf/jclinpath00276-0077e.pdf

5. Center for Disease Control and Prevention - U.S. Department of Health & Human Services. About parasites [Internet]. 2016. Available from: http://www.cdc.gov/parasites/about.html

6. World Health Organization - Infectious diseases [Internet]. [cited 2018 Mar 11]. Available from: http://www.who.int/topics/infectious_diseases/en/

7. World Health Organization - Zoonoses [Internet]. [cited 2018 Mar 11]. Available from: http://www.who.int/zoonoses/diseases/en/

8. Burri C, Keiser J. Pharmacokinetic investigations in patients from northern angola refractory to melarsoprol treatment. Trop Med Int Heal. 2001;6(5):412–20.

9. Sobel JD, Nagappan V, Nyirjesy P. Metronidazole-resistant vaginal trichomoniasis - an emerging problem. N Engl J Med [Internet]. 1999;341(4):292–3. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10419394

10. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet [Internet]. 2012 Dec;380(9859):2095–128. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0140673612617280

11. Krauss H, Weber A, Appel M, Isenberg HD, Schiefer HG, Slenczka W, et al. Zoonoses: Infectious Diseases Transmissible from Animals to Humans. 3rd Edition. [Internet]. American Society of Microbiology Press. 2003. Available from: https://academic.oup.com/cid/article-lookup/doi/10.1093/cid/ciw234

12. Brayton KA, Lau AOT, Herndon DR, Hannick L, Kappmeyer LS, Berens SJ, et al. Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog. 2007;3(10):1401–13.

13. Elmore SA, Jones JL, Conrad PA, Patton S, Lindsay DS, Dubey JP. Toxoplasma gondii: epidemiology, feline clinical aspects, and prevention. Trends Parasitol [Internet]. 2010;26(4):190–6. Available from: http://dx.doi.org/10.1016/j.pt.2010.01.009

14. Pain A, Renauld H, Berriman M, Murphy L, Yeats C a, Weir W, et al. Genome of the host-cell transforming parasite Theileria annulata compared with T. parva. Science [Internet]. 2005;309(5731):131–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15994557

15. Widmer G, London E, Zhang L, Ge G, Tzipori S, Carlton JM, et al. Preliminary Analysis of the Cryptosporidium muris Genome. In: Giardia and Cryptosporidium: from molecules to disease. 2009. p. 320–7.

16. Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, et al. Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. Curr Biol [Internet]. 2016;26(2):161–72. Available from: http://dx.doi.org/10.1016/j.cub.2015.11.055

17. Gabaldón T, Koonin E V. Functional and evolutionary implications of gene orthology. Nat Rev Genet [Internet]. 2013 May 4;14(5):360–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28299348

18. Ncbi. Genomes of protozoa in RefSeq [Internet]. ncbi. 2018 [cited 2018 Mar 3]. Available from: ftp.ncbi.nlm.nih.gov/genomes/refseq/protozoa

19. Koonin E V, Galperin MY. Sequence - Evolution - Function Computational Approaches in Comparative Genomics [Internet]. Boston: Kluwer Academic; 2003. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20260/

20. Fermino BR, Viola LB, Paiva F, Garcia HA, De Paula CD, Botero-Arias R, et al. The phylogeography of trypanosomes from South American alligatorids and African crocodilids is consistent with the geological history of South American river basins and the transoceanic dispersal of Crocodylus at the Miocene. Parasites and Vectors [Internet]. 2013;6(1):1. Available from: Parasites & Vectors

21. Kelly S, Ivens A, Manna PT, Gibson W, Field MC. A draft genome for the African crocodilian trypanosome Trypanosoma grayi. Sci Data [Internet]. 2014;1:1–7. Available from: http://www.nature.com/articles/sdata201424

22. Ivens AC, Peacock CS, Worthey EA, Murphy L, Berriman M, Sisk E, et al. The Genome of the Kinetoplastid Parasite, Leishmania major Alasdair. Science (80- ). 2005;309(5733):436–42.

23. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, et al. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. Science (80- ). 2005;309(5733):409–15.

24. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu D. The Genome of the African Trypanosome Trypanosoma brucei. Science (80- ) [Internet]. 2005;309(5733):416–22. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1112642

25. Jackson AP, Berry A, Aslett M, Allinson HC, Burton P, Vavrova-Anderson J, et al. Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. Pnas [Internet]. 2012;109(9):3416–21. Available from: http://dx.doi.org/10.1073/pnas.1117313109

26. Jackson AP, Allison HC, Barry JD, Field MC, Hertz-Fowler C, Berriman M. A Cell-surface Phylome for African Trypanosomes. PLoS Negl Trop Dis. 2013;7(3).

27. Sistrom M, Evans B, Bjornson R, Gibson W, Balmer O, Maser P, et al. Comparative genomics reveals multiple genetic backgrounds of human pathogenicity in the trypanosoma brucei complex. Genome Biol Evol. 2014;6(10):2811–9.

28. Lumadue JA, Manabe YC, Moore RD, Belitsos PC, Sears CL, Clark DP. A clinicopathologic analysis of AIDS-related cryptosporidiosis. Aids. 1998;12(May):2459–66.

29. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, et al. The genome of Cryptosporidium hominis. Nature. 2004;431(October):557–61.

30. Tyzzer EE. A sporozoan found in the peptic glands of the common mouse. Proc Soc Exp Biol Med. 1907;5:12–3.

31. Nime F.A., Burek J.D., Page D.L., Holscher M.A. YJH. Acute enterocolitis in a human being infected with the protozoan Cryptosporidium. Gastroenterology [Internet]. 1976;70(4):592–8. Available from: http://dx.doi.org/10.1016/S0016-5085(76)80503-3

32. Guyot K, Follet-Dumoulin A, Lelievre E, Sarfati C, Rabodonirina M, Nevez G, et al. Molecular Characterization of Cryptosporidium Isolates Obtained from Humans in France. J Egypt Soc Parasitol. 2001;39(10):3472–80.

33. Tiangtip R, Jongwutiwes S. Molecular analysis of Cryptosporidium species isolated from HIV-infected patients in Thailand. Trop Med Int Heal. 2002;7(4):357–64.

34. Palmer CJ, Xiao L, Terashima A, Guerra H, Gotuzzo E, Saldías G, et al. Cryptosporidium muris, a rodent pathogen, recovered from a human in Per?? Emerg Infect Dis. 2003;9(9):1174–6.

35. NCBI. Cryptosporidium muris - NCBI - Genomes [Internet]. 2018 [cited 2018 Mar 13]. Available from: https://www.ncbi.nlm.nih.gov/genome/?term=Cryptosporidium+muris

36. Xiao L, Limor J, Morgan UM, Sulaiman IM, Thompson RCA, Lal AA. Sequence differences in the diagnostic target region of the oocyst wall protein gene of Cryptosporidium parasites. Appl Environ Microbiol. 2000;66(12):5499–502.

37. Ehrenkaufer GM, Weedall GD, Williams D, Lorenzi HA, Caler E, Hall N, et al. The genome and transcriptome of the enteric parasite Entamoeba invadens, a model for encystation. Genome Biol [Internet]. 2013;14(7):R77. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-7-r77

38. Sanchez L, Enea V, Eichinger D. Identification of a developmentally regulated transcript expressed during encystation of Entamoeba invadens. Mol Biochem Parasitol. 1994;67(1):125–35.

39. Bradford CM, Denver MC, Cranfield MR. Development of a polymerase chain reaction test for Entamoeba invadens. J Zoo Wildl Med. 2008;39(2):201–7.

40. Koonin E V. Orthologs, Paralogs, and Evolutionary Genomics. Annu Rev Genet [Internet]. 2005;39(1):309–38. Available from: http://www.annualreviews.org/doi/10.1146/annurev.genet.39.073003.114725

41. Hardison RC. Comparative genomics. PLoS Biol. 2003;1(2):156–60.

42. Koonin E V, Galperin. MY. Sequence - Evolution - Function Computational Approaches in Comparative Genomics [Internet]. 2003. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20260/

43. Margelevicius M, Venclovas C. Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. BMC Bioinformatics. 2010;11:89.

44. Eisen JA, Wu M. Phylogenetic Analysis and Gene Functional Predictions: Phylogenomics in Action. Theor Popul Biol. 2002;61(4):481–7.

45. Pearson WR. An Introduction to Sequence Similarity (“Homology”) Searching.

Curr Protoc Bioinforma [Internet]. 2013;Chapter 3: Available from: https://www.ncbi.nlm.nih.gov/pubmed/23749753

46. William R.Pearson. Rapid and sensitive sequence comparison with FASTP and. In: Methods in Enzymology. 1990. p. 63–98.

47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990;215(3):403–10. Available from: http://www.sciencedirect.com/science/article/pii/S0022283605803602

48. Remm M, Storm CEV, Sonnhammer ELL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 2001;314(5):1041–52.

49. da Cruz SMS, Mattoso M, Batista V, Dávila AMR, Silva E, Tosta F, et al. OrthoSearch: a scientific workflow approach to detect distant homologies on protozoans. Proc 2008 ACM Symp Appl Comput - SAC ’08 [Internet]. 2008;18:1282. Available from: http://portal.acm.org/citation.cfm?doid=1363686.1363983

50. Li L, Stoeckert CJJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research. Genome Res [Internet]. 2003;13(9):2178–89. Available from: http://genome.cshlp.org/cgi/content/full/13/9/2178

51. Wall DP, DeLuca T. Ortholog Detection Using the Reciprocal Smallest Distance Algorithm. In 2007. p. 95–110. Available from: http://link.springer.com/10.1007/978-1-59745-515-2_7

52. Dessimoz C, Cannarozzi G, Gil M, Margadant D, Roth A, Schneider A, et al. OMA, a comprehensive, automated project for the identification of orthologs

from complete genome data: Introduction and first achievements. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2005;3678 LNBI:61–72.

53. Sonnhammer ELL, Hollich V. Scoredist: a simple and robust protein sequence distance estimator. BMC Bioinformatics [Internet]. 2005;6(1):108. Available from: http://www.biomedcentral.com/1471-2105/6/108

54. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2015;43(D1):D222–6.

55. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. Vol. 5 suppl. 3, Atlas of protein sequence and structure. 1978. p. 345–51.

56. Wheeler TJ, Eddy SR. Nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013;29(19):2487–9.

57. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9(2):173–5.

58. Hughey R, Krogh A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Bioinformatics [Internet]. 1996;12(2):95–107. Available from: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/12/2/95

59. Qian B, Goldstein RA. Performance of an iterated T-HMM for homology detection. Bioinformatics. 2004;20(14):2175–80.

60. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol [Internet]. 2001;313(4):903–19. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0022283601950806

61. Madera M, Gough J. A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res. 2002;30(19):4321–8.

62. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol. 1998;284(4):1201–10.

63. Wistrand M, Sonnhammer EL. Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. BMC Bioinformatics. 2005;6:99.

64. Stormo GD, Fields DS. Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci [Internet]. 1998 Mar;23(3):109–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9581503

65. Sinha S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. Bioinformatics [Internet]. 2006 Jul 15;22(14):e454–63. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl227

66. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, et al. SUPERFAMILY - Sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 2009;37(SUPPL. 1):380–6.

67. Brown T. Genomes. 2nd ed. Oxford: BIOS Scientific Publishers Ltd; 2002.

68. Silva LL, Marcet-Houben M, Nahum LA, Zerlotini A, Gabaldón T, Oliveira G. The Schistosoma mansoni phylome: Using evolutionary genomics to gain insight into a parasite's biology. BMC Genomics. 2012;13(1).

69. Gabaldón T. Large-scale assignment of orthology: back to phylogenetics? Genome Biol [Internet]. 2008;9(10):235. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-10-235

70. Cuadrat RRC, da Serra Cruz SM, Tschoeke DA, Silva E, Tosta F, Jucá H, et al. An orthology-based analysis of pathogenic protozoa impacting global health: an improved comparative genomics approach with prokaryotes and model eukaryote orthologs. OMICS [Internet]. 2014;18(8):524–38. Available from: http://online.liebertpub.com/doi/full/10.1089/omi.2013.0172

71. Ogata H, Goto S, Fujibuchi W, Kanehisa M. Computation with the KEGG pathway database. BioSystems. 1998;47(1–2):119–28.

72. Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, et al. EggNOG v4.0: Nested orthology inference across 3686 organisms. Nucleic Acids Res. 2014;42(D1):231–9.

73. Tatusov RL. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res [Internet]. 2000;28(1):33–6. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.33

74. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, et al. eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res. 2012;40(D1):284–9.

75. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene

ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet [Internet]. 2000 May;25(1):25–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10802651

76. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward Automatic Reconstruction of a Highly Resolved Tree of Life Francesca. Science (80- ). 2006;311(March):1283–8.

77. Chen F. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res [Internet]. 2006;34(90001):D363–8. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkj123

78. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res [Internet]. 2002;30(7):1575–84. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=101833&tool=pmcentrez&rendertype=abstract%5Cnciteulike-article-id:159967%5Cnhttp://dx.doi.org/10.1093/nar/30.7.1575

79. Hubbard TJP, Murzin AG, Brenner SE, Chothia C. SCOP : a Structural Classification of Proteins database. 1997;25(1):236–9.

80. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–85.

81. Baldauf S. The Deep Roots of Eucaryotes. Nature. 2003;300(9):1703–6.

82. Adhianto L, Banerjee S, Fagan M, Krentel M, Marin G, Mellor-Crummey J, et al. HPCTOOLKIT: Tools for performance analysis of optimized parallel

programs. Concurr Comput Pract Exp. 2010;22(6):685–701.

83. Bernardes JS, Vieira FRJ, Zaverucha G, Carbone A. A multi-objective optimization approach accurately resolves protein domain architectures. Bioinformatics. 2015;32(3):345–53.

84. Bernardes J, Zaverucha G, Vaquero C, Carbone A. Improvement in Protein Domain Identification Is Reached by Breaking Consensus, with the Agreement of Many Profiles and Domain Co-occurrence. PLoS Comput Biol. 2016;12(7):1–39.

85. Roth AC, Gonnet GH, Dessimoz C. Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics [Internet]. 2008;9(1):518. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-518

86. Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, et al. Orthology prediction methods: A quality assessment using curated protein families. BioEssays. 2011;33(10):769–80.

87. Forslund K, Pekkari I, Sonnhammer EL. Domain architecture conservation in orthologs. BMC Bioinformatics [Internet]. 2011;12(1):326. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-326

88. Davidson JN, Chen KC, Jamison RS, Musmanno LA, Kern CB. The evolutionary history of the first three enzymes in pyrimidine biosynthesis [Internet]. Vol. 15, BioEssays. 1993. p. 157–64. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8098212

89. Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. Modules, multidomain proteins and organismic complexity. FEBS J. 2005;272(19):5064–78.

90. Asai DJ, Wilkes DE. The Dynein Heavy Chain Family'. 2004;1:23–9.

91. Manning G, Plowman GD, Hunter T, Sudarsanam S. Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci. 2002;27(10):514–20.

92. Fukuchi S, Nishikawa K. Estimation of the number of authentic orphan genes in bacterial genomes. DNA Res. 2004;11(4):219–31.

93. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.

94. Tschoeke DA, Nunes GL, Jardim R, Lima J, Dumaresq ASR, Gomes MR, et al. The comparative genomics and phylogenomics of Leishmania amazonensis parasite. Evol Bioinforma. 2014;10(CI):131–53.

95. Galperin MY, Makarova KS, Wolf YI, Koonin E V. Expanded Microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43(D1):D261–9.

96. Galperin MY, Koonin E V. From complete genome sequence to "complete" understanding? Trends Biotechnol. 2010;28(8):398–406.

97. Motta LS, Da Silva WS, Oliveira DMP, De Souza W, Machado EA. A new model for proton pumping in animal cells: The role of pyrophosphate. Insect Biochem Mol Biol. 2004;34(1):19–27.

98. Ko KM, Lee W, Yu J-R, Ahnn J. PYP-1, inorganic pyrophosphatase, is required for larval development and intestinal function in C. elegans. FEBS Lett [Internet]. 2007;581(28):5445–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17981157

99. Drozak J, Piecuch M, Poleszak O, Kozlowski P, Chrobok L, Baelde HJ, et al. UPF0586 protein C9orf41 homolog is anserine-producing methyltransferase. J Biol Chem. 2015;290(28):17190–205.

100. Quinn PJ, Boldyrev AA, Formazuyk VE. Carnosine: its properties, functions and potential therapeutic applications. Mol Aspects Med. 1992;13:379–444.

101. Poliak P, Hoewyk D Van, Oborník M, Zíková A, Stuart KD, Tachezy J, et al. Functions and cellular localization of cysteine desulfurase and selenocysteine lyase in Trypanosoma brucei. NIH Public Access. 2010;18(7):1089–98.

102. Marijanovic Z, Laubner D, Möller G, Gege C, Husen B, Adamski J, et al. Closing the Gap: Identification of Human 3-Ketosteroid Reductase, the Last Unknown Enzyme of Mammalian Cholesterol Biosynthesis. Mol Endocrinol [Internet]. 2003;17(9):1715–25. Available from: https://academic.oup.com/mend/article-lookup/doi/10.1210/me.2002-0436

103. Beaudoin F, Wu X, Li F, Haslam RP, Markham JE, Zheng H, et al. Functional Characterization of the Arabidopsis -Ketoacyl-Coenzyme A Reductase Candidates of the Fatty Acid Elongase. Plant Physiol [Internet]. 2009;150(3):1174–91. Available from: http://www.plantphysiol.org/cgi/doi/10.1104/pp.109.137497

104. Beltrame-Botelho IT, Talavera-López C, Andersson B, Grisard EC, Stoco PH. A comparative In Silico Study of the antioxidant defense gene repertoire of distinct lifestyle trypanosomatid species. Evol Bioinforma. 2016;12:263–75.

105. KEGG. REACTION : R00897 [Internet]. 2015 [cited 2018 Jun 2]. Available from: http://www.genome.jp/dbget-bin/www_bget?rn:R00897

106. KEGG. REACTION: R00891 [Internet]. 2015 [cited 2018 Jun 2]. Available from:

http://www.genome.jp/dbget-bin/www_bget?rn:R00891

107. Breitinger U, Clausen T, Ehlert S, Huber R, Laber B, Schmidt F, et al. The three-dimensional structure of cystathionine beta-lyase from Arabidopsis and its substrate specificity. Plant Physiol [Internet]. 2001;126(2):631–42. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=111155&tool=pmcentrez&rendertype=abstract

108. Pais FS-M, Ruy P de C, Oliveira G, Coimbra RS. Assessing the efficiency of multiple sequence alignment programs. Algorithms Mol Biol [Internet]. 2014;9:4. Available from: http://dx.doi.org/10.1186/1748-7188-9-4%5Cnhttp://1.10.213.84

109. Tang S-N, Huang J-F. Evolution of different oligomeric glycyl-tRNA synthetases. FEBS Lett [Internet]. 2005 Feb 28;579(6):1441–5. Available from: http://doi.wiley.com/10.1016/j.febslet.2005.01.045

110. Smith TF, Hartman H. The evolution of Class II Aminoacyl-tRNA synthetases and the first code. FEBS Lett [Internet]. 2015;589(23):3499–507. Available from: http://dx.doi.org/10.1016/j.febslet.2015.10.006

111. Hatzfeld M. The armadillo family of structural proteins. Int Rev Cytol [Internet]. 1999;186:179–224. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9770300

112. Coates JC. Armadillo repeat proteins: beyond the animal kingdom. Trends Cell Biol [Internet]. 2003 Sep;13(9):463–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12946625

113. Cozier GE, Carlton J, Bouyoucef D, Cullen PJ. Membrane targeting by

pleckstrin homology domains. Curr Top Microbiol Immunol [Internet]. 2004;282:49–88. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14594214

114. NIH. platelet protein P47 [Internet]. U.S. National Library of Medicine. 1983 [cited 2018 Jun 13]. Available from: https://meshb.nlm.nih.gov/record/ui?name=pleckstrin

115. Eck MJ, Dhe-Paganon S, Trüb T, Nolte RT, Shoelson SE. Structure of the IRS-1 PTB domain bound to the juxtamembrane region of the insulin receptor. Cell [Internet]. 1996 May 31;85(5):695–705. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8646778

116. Krauss G. Biochemistry of Signal Transduction and Regulation [Internet]. Krauss G, editor. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA; 2014. 235 p. Available from: http://doi.wiley.com/10.1002/9783527667475

117. Brannigan JA, Dodson G, Duggleby HJ, Moody PC, Smith JL, Tomchick DR, et al. A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. Nature [Internet]. 1995 Nov 23;378(6555):416–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7477383

118. Oinonen C, Rouvinen J. Structural comparison of Ntn-hydrolases. Protein Sci [Internet]. 2000;9(12):2329–37. Available from: http://doi.wiley.com/10.1110/ps.9.12.2329

119. Bompard-Gilles C, Villeret V, Davies GJ, Fanuel L, Joris B, Frère J-M, et al. A new variant of the Ntn hydrolase fold revealed by the crystal structure of l-aminopeptidase d-Ala-esterase/amidase from Ochrobactrum anthropi. Structure [Internet]. 2000 Feb;8(2):153–62. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0969212600000915

120. Gough J, Karplus K, Hughey R, Chothia C. P-loop containing nucleoside triphosphate hydrolases superfamily [Internet]. 2001. Available from: http://supfam.org/SUPERFAMILY/cgi-bin/scop.cgi?sunid=52540

121. Leipe DD, Koonin E V., Aravind L. STAND, a Class of P-Loop NTPases Including Animal and Plant Regulators of Programmed Cell Death: Multiple, Complex Domain Architectures, Unusual Phyletic Patterns, and Evolution by Horizontal Gene Transfer. J Mol Biol [Internet]. 2004 Oct;343(1):1–28. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0022283604010010

122. Gough J, Karplus K, Hughey R, Chothia C. Nucleotidylyl transferase superfamily [Internet]. 2001. Available from: http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/cgi-bin/scop.cgi?sunid=52374

123. Khan MN, Sakata K, Komatsu S. Proteomic analysis of soybean hypocotyl during recovery after flooding stress. J Proteomics [Internet]. 2015 May;121:15–27. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1874391915001049

124. Xu Y, Carr PD, Vasudevan SG, Ollis DL. Structure of the Adenylylation Domain of E. coli Glutamine Synthetase Adenylyl Transferase: Evidence for Gene Duplication and Evolution of a New Active Site. J Mol Biol [Internet]. 2010 Feb;396(3):773–84. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0022283609015162

125. Ullrich TC. Crystal structure of ATP sulfurylase from Saccharomyces cerevisiae, a key enzyme in sulfate activation. EMBO J [Internet]. 2001 Feb 1;20(3):316–29. Available from: http://emboj.embopress.org/cgi/doi/10.1093/emboj/20.3.316

126. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: Superfamily: beta and

beta-prime subunits of DNA dependent RNA-polymerase [Internet]. 2009 [cited 2018 Mar 15]. Available from: http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.f.bg.b.html

127. Neer EJ, Schmidt CJ, Nambudripad R, Smith TF. The ancient regulatory-protein family of WD-repeat proteins. Nature [Internet]. 1994 Sep 22;371(6495):297–300. Available from: http://www.nature.com/doifinder/10.1038/371297a0

128. Suganuma T, Pattenden SG, Workman JL. Diverse functions of WD40 repeat proteins in histone recognition. Genes Dev. 2008;22(10):1265–8.

129. Kusmierczyk AR, Martin J. Assembly of chaperonin complexes. Mol Biotechnol [Internet]. 2001 Oct;19(2):141–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11725484

130. Kubota H, Hynes G, Willison K. The chaperonin containing t-complex polypeptide 1 (TCP-1). Multisubunit machinery assisting in protein folding and assembly in the eukaryotic cytosol. Eur J Biochem [Internet]. 1995 May 15;230(1):3–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7601114

131. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: Superfamily: GroEL equatorial domain-like [Internet]. 2009. Available from: http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.b.cff.b.html

132. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: Superfamily: Protein kinase-like (PK-like) [Internet]. 2009. Available from: http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.e.daa.b.html

133. Engh RA, Bossemeyer D. Structural aspects of protein kinase control—role of conformational flexibility. Pharmacol Ther [Internet]. 2002 Feb;93(2–3):99–111.

Available from: http://linkinghub.elsevier.com/retrieve/pii/S0163725802001808

134. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science [Internet]. 2002;298(5600):1912–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12471243

135. Hanks SK. Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. Genome Biol [Internet]. 2003;4(5):111. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12734000

# 9 APÊNDICES E/OU ANEXOS

## 9.1 Arquivo adicional 1: Algoritmo de reconciliação

### 9.1.1 LEIAME

Comando para executar o supergroups.rb:
ruby supergroups.rb all_orthomcl_3_taxa.out OrthologousGroups.txt

* all_orthomcl_3_taxa.out = saída do OrthoMCL
* OrthologousGroups.txt = saída do OMA
* Os arquivos omaFile.rb, omaType.rb, orthomclFile.rb, orthomclType.rb, OrthologousGroups.txt and all_orthomcl_3_taxa.out devem estar na mesma pasta que o supergroups.rb

* Com o objetivo de remover grupos homólogos contidos em outros grupos, deve-se executar o script "identifies_groups_contained_in_other.py" editando o arquivo para informar a localização dos grupos homólogos gerados.

* Com o objetivo de remover proteínas repetidas contidas em grupos concatenados, deve-se executar o script "removes_repeated_proteins.rb" com o seguinte comando:
for x in `ls`; do ruby removes_repeated_proteins.rb $x ; done

* Para validar por CDD e Pfam os Supergrupos homólogos:
validate_sg_cdd.py
validate_sg_pfam.py

### 9.1.2 supergroups.rb

```
#!/usr/bin/env ruby

require './orthomclFile.rb'
require './orthomclType.rb'
require './omaFile.rb'
require './omaType.rb'

$parserMcl = OrthoMCLFile.new(ARGV[0])
$parserOma = OmaFile.new(ARGV[1])
file = ARGV[2] unless ARGV[2].nil?

$mcl = $parserMcl.mcl
$oma = $parserOma.oma

puts "Preparando ."
# Groups of MCL with 2 and 3 genes equal to those of OMA
iguais = []
print '..'
$mcl.each do |k, v|
        regMcl = OrthoMCLType.new(v)
        pMCL = regMcl.proteins
        if regMcl.genes == 3
                print "."
                $oma.each do |k1, v1|
                        regOma = OmaType.new(v1)
                        pOMA = regOma.proteins
                        iguais << k if pMCL == pOMA
                end
        end
end

iguais2 = []
print '..'
```

```
$mcl.each do |k, v|
	regMcl = OrthoMCLType.new(v)
	pMCL = regMcl.proteins
	if regMcl.genes == 2
		print "."
		$oma.each do |k1, v1|
			regOma = OmaType.new(v1)
			pOMA = regOma.proteins
			if regOma.genes == 2
				iguais2 << k if pMCL == pOMA
			end
		end
	end
end


def find(type, id)
	puts "Find id: #{id} and Type #{type}"
	if id.nil?
		puts "Terminou recursividade"
		return id
	else
		if type == 1 # Search in OMA
			puts "Buscando no OMA"
			# Grab MCL Proteins
			regMCL                                                    =
OrthoMCLType.new($parserMcl.getProteinsByGroupId(id))
			pMCL = regMCL.proteins
			$oma.each do |k1, v1|
				regOma = OmaType.new(v1)
				pOMA = regOma.proteins
				if (pMCL & pOMA) != []
					puts "Encontrou no OMA, indo para o find MCL"
					unless $sg.include?(k1)
```

```
                                        $sg << k1
                                        find(2, k1)
                                    end
                                end
                            end
                            puts "Nao encontrou no OMA"
                            return nil
                        elsif type == 2   # BSearch in MCL
                            puts "Buscando no MCL"
                            # Grab OMA Proteins
                            regOMA                                                  =
OmaType.new($parserOma.getProteinsByGroupId(id))
                            pOMA = regOMA.proteins
                            $mcl.each do |k1, v1|
                                regMcl = OrthoMCLType.new(v1)
                                pMCL = regMcl.proteins
                                if (pOMA & pMCL) != []
                                    puts "Encontrou no MCL, indo para o find OMA"
                                    unless $sg.include?(k1)
                                        $sg << k1
                                        find(1, k1)
                                    end
                                end
                            end
                            puts "Nao encontrou no MCL"
                            return nil
                        end
                end
end


# Groups with some intersection without equals (2 and 3 genes)
puts "Criando Super-Grupos MCL x OMA ."
print '..'
```

```
usados = iguais + iguais2
Kernel.system("echo #{usados} > iguais.txt")
$mcl.each do |k, v|
        regMcl = OrthoMCLType.new(v)
        pMCL = regMcl.proteins
        print "."
        if not usados.include?(k)
                $sg = []
                $sg << k
                achou = find(1, k)
                puts "#{k}: #{$sg.inspect}" if $sg.size > 1
                #Kernel.system("echo #{k}\t#{$sg.inspect} > SG_#{k}") if $sg.size > 1

        end

end
```

### 9.1.3 orthomclFile.rb

```
        class OrthoMCLFile

           attr_accessor :mcl

           def initialize(file)
              mcl = File.open(file).readlines

              # Creates a HASH for the result
              mclH = {}

              # reads the file, line by line
              mcl.each do |line|

                 # Get the ID of the orthographic group
                 id = line.split(':')[0].split('(')[0]
```

```
            # Creates an array for the group's proteins
            mclH[id] = line.split(':')[1].split(" ")
        end
        @mcl = mclH
    end

    def inspect
        puts @mcl.inspect
    end

    # Get Proteins By Group Id
    def getProteinsByGroupId(id)
        return @mcl[id]
    end

end
```

### 9.1.4 orthomclType.rb

```
class OrthoMCLType

    attr_accessor :type, :genes, :proteins, :taxa, :organisms

    def initialize(type)
        if type.class == Array
            @type = type
        else
            raise 'Type wrong!'
        end
    end

    # Number of proteins in group
    def genes
```

```
        return @type.size
    end

    # Number of Taxas
    def taxa
        tmp = []
        @type.each do |p|
            tmp << p.split('(')[1].split(')')[0]
        end
        tmp.sort!
        tmp.uniq!
        return tmp.size
    end

    # Array of Organisms
    def organisms
        tmp = []
        @type.each do |p|
            tmp << p.split('(')[1].split(')')[0]
        end
        return tmp
    end
    # Array of Proteins ID
    def proteins
        tmp = []
        @type.each do |p|
            tmp << p.split('(')[0]
        end
        return tmp
    end

end
```

### 9.1.5 omaFile.rb

```
class OmaFile

   attr_accessor :oma

   def initialize(file)
      oma = File.open(file).readlines

      # Creates a HASH for the result
      omaH = {}

      # reads the file, line by line
      oma.each do |line|
         # Ignore lines that begin with the #
         unless line =~ /^#/
            x = line.split("\t")
            id = x[0]
            tmp = []
            i = 1
            x.each do |p|
               unless i == 1
                  tmp << p.gsub("\n",'')
               end
               i+=1
            end
            omaH[id] = tmp
         end
      end
      @oma = omaH
   end

   def inspect
```

```
      puts @oma.inspect
    end

    # Get Proteins By Group Id
    def getProteinsByGroupId(id)
      return @oma[id]
    end

end
```

### 9.1.6 omaType

```
        class OmaType

            attr_accessor :type, :genes, :proteins, :taxa, :organisms

            def initialize(type)
                if type.class == Array
                    @type = type
                else
                    raise 'Type wrong!'
                end
            end

            # Number of proteins in group
            def genes
                return @type.size
            end

            # Number of Taxas
            def taxa
                tmp = []
                @type.each do |p|
                    tmp << p.split(':')[0]
```

```
        end
        tmp.sort!
        tmp.uniq!
        return tmp.size
    end

    # Array of Organisms
    def organisms
        tmp = []
        @type.each do |p|
            tmp << p.split(':')[0]
        end
        return tmp
    end

    # Array of Proteins ID
    def proteins
        tmp = []
        @type.each do |p|
            tmp << p.split(':')[1].split(' ')[0]
        end
        return tmp
    end

end
```

### 9.1.7 identifies_groups_contained_in_other.py

```
#!/usr/bin/python3
# -*- coding: utf-8 -*-

from glob import glob

sg = '/home/darueck1/artigo/SG/'
```

```
omcl = '/home/darueck1/artigo/OMCL/'

acr = []

def scount(fl):
    handler = open(fl, 'r')
    text = handler.read().strip()
    handler.close()
    return(int(text.count('>')))


for fl_path in glob(sg+'*.fa*'):
    sg_c = scount(fl_path)
    sg_name = fl_path.split('/')[-1].strip()
    orth_name = (((sg_name.replace('SG_', 'ORTHOMCL')).split('.')[0])+'.fasta')
    om_c = scount((omcl+orth_name).replace('//','/'))
    print('%s %i\n%s %i\n' % (sg_name, sg_c, orth_name, om_c))
    if sg_c > om_c:
        diff = (sg_c - om_c)
        acr.append('Name:%s\nQTD:%i\nAddition of:%i sequences\nPath:%s' %
(sg_name, sg_c, diff, fl_path))


print('\n\n\n'+'---------\n\nIncreased the number of sequences:\n')

for i in acr:
    print(i)
    print()
```

### 9.1.8 removes_repeated_proteins.rb

```
# Line to execute
# "for x in `ls`; do ruby removes_repeated_proteins.rb $x ; done"


sgname = ARGV[0]
sg = File.open(ARGV[0]).readlines
x = []

sg.each do |line|

        x << line
        x.sort!
   x.uniq!
end

puts x
puts sgname

File.open("PSG_#{sgname}", "w+") do |f|
                                    f.puts(x)
end
```

### 9.1.9 validate_sg_cdd.py

```
#!/usr/bin/python3

from glob import glob

supergrp = {}
CDD = {}


def write(file_path, content):
```

```
    handler = open(file_path, 'a')
    handler.write(content)
    handler.close()




for fasta in glob('/home/darueck1/artigo/SG/*.fa'):
    handler = open(fasta, 'r')
    content = (set(handler.read().strip().split('\n')))
    handler.close()
    temp_seq = list(filter((lambda x: ('>' in x)), content))
    temp_seq = list(map((lambda x: (x.replace('>',''))), temp_seq))
    for seq in temp_seq:
        prot_id = seq.split()[0]
        seq = (seq.replace(prot_id, '')).strip()
        sg = ((fasta.split('/')[-1]).split('.')[0].strip())
        if sg not in supergrp:
            supergrp[sg] = []
            supergrp[sg].append(prot_id)
        else:
            supergrp[sg].append(prot_id)
    handler.close()



handler = open('/home/darueck1/artigo/lista_cdd_prt.txt', 'r')
prt_lista = handler.read().strip().split('\n')
handler.close()

for prt in prt_lista:
    prt_id = prt.split()[0]
    if prt_id not in CDD:
        CDD[prt_id] = prt.split()[-1]
```

```
del prt_lista


write('sg_EQU_CDD.csv', 'SG,PROTEIN,CDD\n')
write('sg_NO_CDD.csv', 'SG,PROTEIN,CDD\n')
write('sg_DIF_CDD.csv', 'SG,PROTEIN,CDD\n')


rst = []

for sg in supergrp.keys():
    for prt in supergrp[sg]:
        rst.append('%s,%s,%s' % (sg, prt, CDD[prt]))


spf = list(filter((lambda x: ('SEM_' in x)), rst))
spf_grp = list(set(map((lambda x: (x.split(',')[0].strip())), spf)))


no_CDD = sorted(set(filter((lambda x: ((x.split(',')[0].strip()) in spf_grp)), rst)))


for item in no_CDD:
    rst.remove(item)



equ_CDD = []

for grp in set(map((lambda x: (x.split(',')[0])), rst)):
    grp_rst = list(set(filter((lambda x: grp in x), rst)))
    if len(set(map((lambda x: x.split(',')[-1]), grp_rst))) == 1:
        equ_CDD.extend(grp_rst)
```

```
equ_CDD = sorted(set(equ_CDD))


for gp in set(map((lambda x: (x.split(',')[0])), equ_CDD)):
    if gp in spf_grp:
        print(gp)


for item in equ_CDD:
    rst.remove(item)


rst = sorted(set(rst))


write('sg_EQU_CDD.csv', ('\n'.join(equ_CDD)))
write('sg_NO_CDD.csv', ('\n'.join(no_CDD)))
write('sg_DIF_CDD.csv', ('\n'.join(rst)))
```

### 9.1.10 validate_sg_pfam.py

```
#!/usr/bin/python3

from glob import glob

supergrp = {}
PFAM = {}


def write(file_path, content):
    handler = open(file_path, 'a')
    handler.write(content)
    handler.close()
```

```
for fasta in glob('/home/darueck1/artigo/SG/*.fa'):
    handler = open(fasta, 'r')
    content = (set(handler.read().strip().split('\n')))
    handler.close()
    temp_seq = list(filter((lambda x: ('>' in x)), content))
    temp_seq = list(map((lambda x: (x.replace('>',''))), temp_seq))
    for seq in temp_seq:
        prot_id = seq.split()[0]
        seq = (seq.replace(prot_id, '')).strip()
        sg = ((fasta.split('/')[-1]).split('.')[0].strip())
        if sg not in supergrp:
            supergrp[sg] = []
            supergrp[sg].append(prot_id)
        else:
            supergrp[sg].append(prot_id)
    handler.close()



handler = open('/home/darueck1/artigo/lista_pfam_prt.txt', 'r')
prt_lista = handler.read().strip().split('\n')
handler.close()

for prt in prt_lista:
    prt_id = prt.split()[0]
    if prt_id not in PFAM:
        PFAM[prt_id] = prt.split()[-1]

del prt_lista
```

```
write('sg_EQU_PFAM.csv', 'SG,PROTEIN,PFAM\n')
write('sg_NO_PFAM.csv', 'SG,PROTEIN,PFAM\n')
write('sg_DIF_PFAM.csv', 'SG,PROTEIN,PFAM\n')


rst = []

for sg in supergrp.keys():
    for prt in supergrp[sg]:
        rst.append('%s,%s,%s' % (sg, prt, PFAM[prt]))


spf = list(filter((lambda x: ('SEM_' in x)), rst))
spf_grp = list(set(map((lambda x: (x.split(',')[0].strip())), spf)))


no_PFAM = sorted(set(filter((lambda x: ((x.split(',')[0].strip()) in spf_grp)), rst)))


for item in no_PFAM:
    rst.remove(item)



equ_PFAM = []

for grp in set(map((lambda x: (x.split(',')[0])), rst)):
    grp_rst = list(set(filter((lambda x: grp in x), rst)))
    if len(set(map((lambda x: x.split(',')[-1]), grp_rst))) == 1:
        equ_PFAM.extend(grp_rst)

equ_PFAM = sorted(set(equ_PFAM))
```

```
for gp in set(map((lambda x: (x.split(',')[0])), equ_PFAM)):
    if gp in spf_grp:
        print(gp)


for item in equ_PFAM:
    rst.remove(item)


rst = sorted(set(rst))


write('sg_EQU_PFAM.csv', ('\n'.join(equ_PFAM)))
write('sg_NO_PFAM.csv', ('\n'.join(no_PFAM)))
write('sg_DIF_PFAM.csv', ('\n'.join(rst)))
```

## 9.2 Arquivo adicional 1: Script Superfamily - bestHits.rb

```
# Executar o script na pasta contendo os resultados do programa COMA
#/usr/bin/env ruby
#command to run (in the folder with the COMA results): "ruby bestHits.rb"#
result = {}
Dir.entries(ARGV[0]).each do |file|
        if file =~ /result/
                f = File.open(file).readlines
                i = 1
                key = ''
                value = ''
                f.each do |line|
                        if line =~ /^lcl\|.{1,}\s.*\s{1,}\d{3,}\s{1,}.{1,}$/
                                key = line.split('|')[1].split(' ')[0] if i == 1
                                if i == 2
                                        value = line.split('|')[1].split(' ')[0]
                                        vec = line.split('|')[1].split(' ')
                                        evalue = vec[vec.size-1].to_f
                                value = '' if evalue > "1E-05".to_f
                                end
                        i+=1
                end
                end
                result[key] = value
        end
end
result2 = result
result.each do |k, v|
        result2.each do |k1, v1|
        puts "#{k} is best hit reciprocal with #{k1}" if (k == v1) and (k1 == v)
        end
end
```