Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz
Instituto Oswaldo Cruz
Curso de Pós-Graduação em Biologia Celular e Molecular

Kary Ann del Carmen Soriano Ocaña

Usando uma abordagem filogenômica para o estudo dos protozoários

Rio de Janeiro
Abril 2010

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ
Pós-Graduação em Biologia Celular e Molecular

Kary Ann del Carmen Soriano Ocaña

Usando uma abordagem filogenômica para o estudo dos protozoários

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Biologia Molecular e Celular

ORIENTADOR: Prof. Dr. Alberto Martín Rivera Dávila

Rio de Janeiro
Abril 2010

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Pós-Graduação em Biologia Celular e Molecular

Kary Ann del Carmen Soriano Ocaña

Usando uma abordagem filogenômica para o estudo dos protozoários

ORIENTADOR: Prof. Dr. Alberto Martín Rivera Dávila

Aprovada em: 27/04/2010

EXAMINADORES:

Prof. Dr. Antonio Basilio de Miranda
Prof. Dr. Carlos Guerra Schrago
Prof. Dr. Christian Macagnan Probst

SUPLENTES:

Prof. Maria Luiza Machado Campos
Prof. Adeilton Alves Brandão

Rio de Janeiro, 27 de abril de 2010

Soriano Ocaña, Kary Ann del Carmen
Usando uma abordagem filogenômica para o estudo dos Protozoários / Kary Ann del Carmen Soriano Ocaña. – Rio de Janeiro: 2010.

Tese (Doutorado) – Instituto Oswaldo Cruz, Biologia Celular e Molecular, 2010.

1. Elementos Genéticos Móveis 2. Protozoários. 3. Filogenômica. 4.Bioinformática I.Título

A Deus por ser minha principal inspiração.

Aos meus queridos vovôs Anita e Miguelito pelo seu carinho infinito.
Os levo no meu coração e os lembro a cada instante e em cada lugar...
Vocês foram os melhores pais do mundo!
Obrigada por tudo.

A minha mãe Gladys e irmãs Joann e Gladys, pela sua alegria, confiança apóio e incomparável carinho. Pelos momentos maravilhosos e os risos incomparáveis. A vocês irmãs por serem as minhas melhores amigas!
Adoro vocês!

A Jhonny pela sua compreensão, disciplina e afeto, a quem considero como um pai.

Em especial as minhas sobrinhas queridas Joannita e Samanthita,
pelos dias lindos, as lembranças doces e meigas, por me ensinarem o simples e fascinante da vida e por simplesmente serem as nenéns mais lindas
Adoro vocês!

A minha família em Peru. Deles aprendi que existe um vínculo que vá, mais além da distância e do tempo.

Ao Olivier pelo amor, compreensão, sensibilidade e amizade. Pela nossa vida maravilhosa no Brasil.
e sobretudo por ser a minha família!

Ao meu orientador Dr. Alberto Martín Rivera Dávila, pelo apoio, confiança e orientação, a quem expresso a minha profunda admiração, respeito e carinho.

À Dra. Yara Traub Cseko, por a sua disposição, atenção, conselhos e amizade.

À Dra. Kimmen Sjölander e ao Dr. Bryan Kolaczkowski do Grupo de Filogenômica de Berkeley, pelas discussões sobre filogenômica e bioinformática.

Ao Dr. Carlos Guerra, por ter aceitado revisar a minha dissertação e participar desta banca examinadora.

Aos doutores Antonio Basílio, Adeilton Alves, Maria Luiza Machado e Christian Probst, por ter aceitado participar desta banca examinadora.

À Ms. Rodrigo Jardim pela amizade e conselhos no âmbito computacional que tanto colaboraram no desenvolvimento deste trabalho.

À Simone Ferreira pelos conselhos na revisão do texto da tese.

Aos colegas do Laboratório de Biologia Computacional e Sistemas e do Laboratório de Biologia Molecular de Tripanossomatídeos e Flebotomíneos pela amizade e apoio.

Aos queridos amigos da Fiocruz: André, Chris, Claudia, Denise, Davi, Edgar, Erich, Erickinha, Felipe, Glauber, Joana, Juliana, Luisa, Mary e Silvana pelos momentos maravilhosos juntos.

Aos amigos e colegas da bioinformática: Adriana, Bernardo, Carol, Carolina, Diogo, Fábio, Marcos, Milene e Rafael pela ajuda, conselhos e momentos divertidos no laboratório.

Aos meus grandes amigos: Ana, Augusto, Bernard, Blanco, Califa, Cid, Flávia, Ivo, Leo, Nelson e Rafael pelo carinho e as reuniões sempre divertidas.

Ao todos os meus amigos e colegas em Peru: da escola, da universidade e da Empresa *Ebel Technological Institute*. Especialmente à Any, Marthita, Rosita e Giovanna que sempre estão presentes com os seus conselhos e carinho.

Ao meu querido tio Dr. Marco Túlio Zapata pelo seu carinho e conselhos.

À Marli, pela amizade sincera e a ternura que me inspira. Por saber e sentir que posso confiar sempre em você!

À Margarita, você é tudo de bom! Obrigada pela amizade, carinho e por estar sempre presente nesses momentos onde uma precisa de uma amiga incondicional.

Ao Dr. Omar Triana pelos conselhos e pela amizade linda.

À CAPES pelo apoio financeiro, e ao Instituto Oswaldo Cruz - FIOCRUZ pela oportunidade e apoio.

Resumo

A reconstrução da história evolutiva, assim como o estabelecimento de hipóteses que demonstrem as relações filogenéticas dos protozoários bem como dos genes codificados pelos Elementos Genéticos Móveis (EGM) requerem o uso de várias abordagens e ferramentas, as quais não se encontram disponíveis de maneira integrada nem de maneira amigável. Diferentes abordagens filogenéticas, filogenômicas e evolutivas são necessárias para a inferência da filogenia de espécies e o estudo de genes pouco conservados como a transcriptase reversa, o gene mais representativo da classe I dos EGM, os retrotransposons. Os principais algoritmos filogenéticos e os programas que os executam têm sido unificados num único sistema: ARPA, escrito na linguagem de programação PYTHON. O sistema ARPA e a interface *web* estão hospedados na FIOCRUZ e estão disponíveis no endereço http://arpa.biowebdb.org. Eles estão sendo integrados ao sistema de banco de dados ProtozoaDB (http://protozoadb.biowebdb.org) e ao sistema de anotação semi-automática Stingray (http://stingray.biowebdb.org/). Uma abordagem baseada nos fundamentos da filogenômica e evolução foi utilizada para desenvolver cinco objetivos: (i) analisar e inferir a filogenia dos genes relacionados à resistência de drogas em protozoários, (ii) reconstruir a árvore de espécies de protozoários, (iii) realizar estudos de filogenômica dos EGM em protozoários, (iv) inferir a filogenia da telomerase e dos elementos de retrotransposição em Tri-tryps e (v) adaptar e ampliar o esquema Phylo ao banco de dados GUS para o armazenamento da informação filogenética.

Os principais resultados obtidos para cada objetivo são: (i) As inferências filogenéticas dos genes AQP, hsp70, GP63, TRYR e MRPA relacionados à resistência a drogas em protozoários demonstrou a viabilidade das execuções do sistema ARPA; (ii) a árvore de espécies de protozoários usando a abordagem da supermatriz provou ser confiável, e o teste PTP e a estatística G1 demonstraram que os dados moleculares deste estudo possuem sinal filogenético; (iii) o RAXML foi o programa mais consistente ao lidar com os diferentes níveis de polimorfismos destes genes, a detecção *in silico* da seleção positiva destes genes foi detectada nas análises pareadas dos modelos M1-M2 e M7-M8, porém o par M0-M3 indicou uma alta variabilidade da razão ω entre os sítios; (iv) foi observada a monofilia para a telomerase a que está mais relacionada à transcriptase reversa dos retrotransposons não-LTR; (iv) um novo esquema Phylo foi concebido e incorporado no GUS 3.5 estendendo-o a fim de armazenar os dados obtidos de inferências filogenéticas.

As principais conclusões são: (i) O sistema ARPA é uma alternativa viável, eficiente, fácil e de tempo reduzido para as análises filogenômicas. O RAXML foi considerado o programa mais consistente e foi observado que as árvores construídas usando as sequências inteiras e/ou as trimadas com o TRIMAL apresentaram os melhores resultados. A abordagem da supermatriz apresentou melhores resultados do que a superárvore; (ii) as relações entre os grupos de protozoários estão de acordo com estudos anteriores da literatura, os quais determinaram também uma monofilia para os protozoários. A inclusão de mais dados/genes é necessária para obter uma árvore robusta; (iii) foram reconstruídas as árvores dos genes dos EGM e inferida a filogenia para cada um deles. O modelo M3 indicou uma alta variabilidade da razão ω entre os sítios e os modelos M7 e M8 indicaram a presença de seleção positiva para todos os genes dos EGM; (iv) a telomerase formou um grupo monofilético mais relacionado à transcriptase reversa dos retrotransposons não-LTR; (v) o esquema Phylo armazena os dados obtidos de experiências filogenéticas, mantendo as relações de herança filogenética entre cada um dos táxons, o que permite realizar consultas usando as informações dos ramos, dos nós e táxons da árvore.

Abstract

The reconstruction of the evolutionary history, as well as the establishment of the hypotheses that demonstrate the phylogenetic relationships of the genes encoded by Mobile Genetic Elements (MGEs) require the use of various tools and approaches, which are not available in a friendly or integrated interface. Different phylogenetics, phylogenomics and evolutionary approaches are necessary for the inference of the species phylogeny. These same approaches are required on the study of less conserved genes as the reverse transcriptase that is the most representative gene of the class I of the MGEs, the retrotransposons. The main phylogenetic algorithms and programs developed by our group have been unified into a single system - the ARPA - written in the programming language PYTHON. The ARPA system and the web interface are hosted at FIOCRUZ and are available at http://arpa.biowebdb.org. They are currently being integrated to the database system ProtozoaDB (http://protozoadb.biowebdb.org) and to the semi-automatic annotation system Stingray (http://stinngray.biowebdb.org/). An approach based on the fundamentals of evolution and phylogenomics has been applied to achieve five different objectives: (i) to analyze and to infer the phylogeny to the genes related to drug resistance in protozoan genomes, (ii) to reconstruct a protozoan species tree, (iii) to conduct phylogenomic studies of MGEs in Protozoa, (iv) to infer phylogeny from the telomerase and the retrotransposable elements in Tri-Tryps and (v) to adapt and to extend the schema Phylo to the GUS database, for storing phylogenetic informations.

The results obtained for topics were: (i) The construction of the phylogenetic trees of the genes, AQP, hsp70, GP63, TRYR and MRPA which are related to drug resistance in protozoan demonstrated the viability of the executions of the ARPA system. (ii) The protozoan species tree using the supermatrix approach proved to be reliable. The PTP Test and the Statistical G1 demonstrated that the molecular data of this study have phylogenetic signal. (iii) The PAUP-AV was shown to be the most consistent program and the PHYML was the least to deal with different levels of polymorphisms of these genes. The *in silico* detection of the positive selection in MGEs genes in Protozoa was detected in the paired analysis of the models M1-M2 and M7-M8, but the pair M0-M3 indicated a high variability of the ratio ω between the sites. (iv) It was found that a monophyly is present for the telomerase, which was the most closely related to the transcriptase of the non-LTR retrotransposons. (v) A new Phylo schema was designed and incorporated into the GUS 3.5 extending its service to store the data obtained from phylogenetic experiments.

As conclusions: (i) The ARPA system is a viable, efficient, easy and reduced time alternative for phylogenomic analysis. The RAXML was considered the most consistent program and was observed that the trees constructed using the entire and/or the trimmed sequences with TRIMAL showed the best results. The supermatrix approach showed better results than the supertree. (ii) The relationships between protozoan groups are in agreement with previous studies, which also determined a monophyly for protozoan. The inclusion of more data/genes is required to obtain a consistent tree. (iii) In the trees of the EGM, the PAUP-AV was the most consistent and the PHYML the least to deal with different levels of polymorphisms of these genes. The model M3 showed a high variability of ω ratio among sites and the models M7 and M8 indicated the presence of positive selection for all genes of EGM. (iv) The telomerase formed a monophyletic group more related to the reverse transcriptase of the non-LTR retrotransposons. (v) The scheme Phylo stores the data obtained from phylogenetic experiences, keeping the inheritance of phylogenetic relationships between each of the taxa, which can perform queries using information from the branches, nodes and taxa of the tree.

Lista de figuras

Lista de tabelas

Lista de anexos

Lista de abreviaturas, siglas e símbolos

BIOLOGIA

DNA	Ácido desoxirribonucléico
RNA	Ácido ribonucléico
mRNA	RNA mensageiro
iRNA	RNA de interferência
cDNA	DNA complementar
kDNA	DNA do cinetoplasto
Kb	Kilobases
pb	Pares de base
ORF	Janela aberta de leitura (do inglês, *open reading frame*)
EST	(do inglês, *expressed sequence target*).

ELEMENTOS GENÉTICOS MÓVEIS

EGM	Elementos Genéticos Móveis
LTR	Repetições Terminais Longas
não-LTR	Repetições Terminais não Longas
DSA	Duplicações de Sítio Alvo (do inglês, *target site duplications TSDs*)
Tn	Transposons Compostos
TR	Transcriptase Reversa
EN	Endonuclease
RH	Ribonuclease H
IN	Integrase
gag	Genes antígenos grupo-específico
prt	Protease
pol	Polimerase
env	Envelope
TERT	Telomerase
GII-I	Grupo do íntron II

HERV	Retrovírus Endógenos Humanos
LINE	Elementos Nucleares Longos Intercalados
SINE	Elementos Nucleares Curtos Intercalados
LINE-1 ou L1	LINE em humanos
NAR*Tc*	Retrotransposon não-autônomo em *T. cruzi* (do inglês, *non-autonomous retrotransposon in T. cruzi*)
L1Tc	LINE em *T. cruzi*
SIRE	Elementos Repetitivos Curtos Intercalados (do inglês, *short interspersed repetitive elements*)
VIPER	Vestígios de retroelementos Intercalados (do inglês, *vestigial interposed retroelements*)
TART	Retrotransposon Associado ao Telômero
RLMas	Retrotransposons LTR de mamíferos

BIOINFORMÁTICA

BLAST	Pacote (do inglês, *Basic Local Alignment Search Tool*)
CPU	Unidad Central de Procesamiento (do inglês, *Central Processing Unit*)
HMM	Modelo estatístico (do inglês, *hidden Markov models*)
HMMER	Pacote, é uma implementação de *software* distribuído gratuitamente do perfil HMM para a análise da sequência de proteínas
hmmbuild	Programa do pacote HMMER
hmmcalibrate	Programa do pacote HMMER
hmmpfam	Programa do pacote HMMER
perfis HMM	(do inglês, *profiles*) HMM
PERL	Linguagem de programação (do inglês, *Practical Extraction and Report Language*)
BIOPERL	Librarias de *scripts* em módulos para biologia (do inglês, *Perl scripts for bioinformatics applications*)
HTML	Linguagem de programação para a criação de páginas *web* (do inglês, *HyperText Markup Language*)

SWISS-PROT	Banco de dados biológicos curados de sequências protéicas
NCBI	Banco de dados (do inglês, *National Center for Biotechnology Information*)
PFAM	Banco de dados (do inglês, *Protein families database - HMM derived*)

FILOGENIA

AV	Agrupamento de vizinhos (do inglês, *neighbor-joining*)
ME	Evolução mínima (do inglês, *minimum-evolution*)
MV	Máxima verossimilhança (do inglês, *maximum likelihood*)
IB	Inferência Bayesiana (do inglês, *Bayesian inference*)
MP	Máxima parcimônia (do inglês, *maximum parsimony*)
PAML	Pacote (do inglês, *Phylogenetic Analysis by Maximum Likelihood*)
CODEML	Programa parte do pacote PAML
PHYLIP	Pacote (do inglês, *PHYLogeny Inference Package*)
PAUP	Programa (do inglês, *Phylogenetic Analysis Using Parsimony*)
PHYML	Programa (do inglês, *PHYlogenies by Maximum Likelihood*)
WEIGHBOR	Programa (do inglês, *Weighted Neighbor Joining*)
TREE-PUZZLE	Programa
RAXML	Programa (do inglês, *Randomized Axelerated Maximum Likelihood*)
GARLI	Programa (do inglês, *Genetic Algorithm for Rapid Likelihood Inference*)
MRBAYES	Programa (do inglês, *Bayesian Analysis of Phylogeny*)

CAPÍTULO 1 - INTRODUÇÃO

1.1 Reconstrução automática de análises filogenômicas (ARPA) de genes relacionados à resistência a drogas em genomas de protozoários

1.1.1 Os genes de resistência a drogas em protozoários

Doenças causadas por parasitas protozoários são uma das principais causas de mortalidade e morbidade em humanos e animais domésticos em todo o mundo. Essas doenças estão concentradas nas regiões mais pobres e recebem relativamente pouca atenção da indústria farmacêutica e do mercado.

O controle das principais infecções parasitárias é feito através da quimioterapia, pois suas vacinas ainda estão em desenvolvimento. No entanto, várias doenças graves (por exemplo: malária, leishmaniose e tripanossomíase), bem como outras menos perigosas (por exemplo: amebíase e tricomoníase), causadas por protozoários parasitas, continuam sendo responsáveis por um aumento alarmante de casos refratários ao tratamento (Arango *et al*. 2008; Arevalo *et al*. 2007; Bansal *et al*. 2004; Burri *et al*. 2001; Sobel *et al*. 1999).

O fracasso do tratamento tem, potencialmente uma origem multifatorial, mas uma das maiores preocupações é a de resistência às drogas. Esta situação levou ao início de vários projetos genoma, com o objetivo de esclarecer a bioquímica básica dos parasitas mais perigosos e acelerar a identificação dos alvos para novas drogas. Isto tem requerido muito investimento, tanto financeiro como de trabalho e tempo. Assim, várias abordagens de bioinformática têm sido propostas para minimizar ou contornar algumas das dificuldades encontradas na descoberta de novas drogas antiparasitárias, com qualidades similares ou superiores às atuais (Agrafiotis *et al*. 2007; Myler 2008) baseada na análise filogenética e na predição de genes de resistência a drogas (Fuellen *et al*. 2005).

Existem vários mecanismos de resistência já identificados em microorganismos parasitas. Apesar do fato de parasitas intracelulares sobreviverem em um ambiente intenso de oxidação, eles também tem de superar os efeitos das drogas para manterem-se vivos. Portanto, vários mecanismos deveriam trabalhar em sinergia para atacar os seus mecanismos de resistência, por

exemplo, aumentando as condições ambientais extremas, bloqueando o transporte das membranas e reduzindo a carga viral (Ashutosh *et al*. 2007; Delespaux *et al*. 2007; Fidock *et al*. 2008).

O presente estudo demonstra a viabilidade das execuções do ARPA na construção das árvores filogenéticas de cinco genes relacionados à resistência a drogas em protozoários:

(i) aquaporina (AQP), transportador de membrana;

(ii) *heat-shock protein* 70 (hsp70), relacionado à proteção ao estresse;

(iii) glicoproteína de 63 kDa (GP63), relacionado à virulência;

(iv) tripanotiona redutase (TRYR), relacionado à destoxificação;

(v) ABC-associado à proteína A (MRPA), transportador de membrana.

Protozoários parasitas são responsáveis por um amplo espectro de doenças em humanos e animais domésticos. A principal linha de defesa disponível contra estes organismos é a quimioterapia. No entanto, a aplicação de quimioterápicos resultou no desenvolvimento de mecanismos de resistência, o que limita o número de medicamentos antiprotozoários eficazes no tratamento e controle de doenças parasitárias. O conhecimento sobre os mecanismos de resistência envolvidos pode levar ao desenvolvimento de novas drogas que minimizem esta resistência.

O controle da leishmaniose é feito apenas através da quimioterapia, pois vacinas contra este parasita ainda estão em desenvolvimento (Brandonisio *et al*. 2002). Antimoniais pentavalentes e seus derivados vêm sendo utilizados como a primeira linha de drogas, mas doenças como a leishmaniose estão se propagando devido a mudanças no ambiente, a imunidade do hospedeiro e a falhas terapêuticas (Croft *et al*. 2006a; Croft *et al*. 2006b; Palatnik-de-Sousa 2008; Sauvage *et al*. 2009; Sundar *et al*. 2000; Yardley *et al*. 2006).

Os componentes mais comuns da interface parasita-hospedeiro são as proteínas da família aquaporina. Estes pequenos canais de membrana integrais são específicos para a água (aquaporinas ortodoxas) ou deixam passar, preferencialmente, solutos polares sem carga, como o glicerol e a uréia (aquagliceroporina) (Engel 2000). A partir das análises dos genomas, tornou-se evidente que parasitas pertencentes ao filo Apicomplexa expressam uma forte redução de aquaporinas quando comparados com os pertencentes ao filo Kinetoplastida (Gourbal *et al*. 2004; Hansen *et al*. 2002; Pavlovic-Djuranovic *et al*. 2003; Uzcategui *et al*. 2004). Dentro dos Apicomplexa, as espécies de *Plasmodium* analisadas até agora - *P. falciparum*, *P. berghei*, *P.*

yoelii, *P. chabaudi*, *P. knowlesi* -, assim como as espécies *T. gondii* e *E. tenella* carregam um gene aquaporina único. Até cinco genes aquaporinas foram identificados nos genomas dos Kinetoplastida *T. brucei*, *T. cruzi* e *L. major*, enquanto que na espécie apicomplexa *C. parvum* não foi identificado um único gene aquaporina. Até hoje, esta é a primeira identificação de um organismo eucariótico que carece totalmente de aquaporinas. O genoma do Apicomplexa *Cryptosporidium* é de aproximadamente 9 Mb (Abrahamsen *et al*. 2004), ou seja, menos da metade do tamanho das espécies do gênero *Plasmodium* (23 Mb)(Gardner *et al*. 2002) ou cerca de um décimo do tamanho da espécie *T. gondii* (87 Mb) (Ajioka *et al*. 2001). O número de genes preditos também é muito menor: aproximadamente 3.800, se comparado aos mais de 5.200 em *Plasmodium* e mais de 6.600 em *Toxoplasma* (Decuypere *et al*. 2005; Huang *et al*. 2004; Pavlovic-Djuranovic *et al*. 2003).

Durante a evolução, as células teriam desenvolvido mecanismos de prevenção e controle de qualidade para evitar a agregação de proteínas. Um deles aconteceria através da síntese das chaperonas moleculares. Dentre as várias proteínas desta categoria, destaca-se o sistema hsp70, que desempenha funções essenciais no metabolismo de proteínas, pois age como um pivô, recebendo e distribuindo proteínas desenoveladas ou substratos entre as demais chaperonas moleculares. Assim, o estudo do mecanismo de ação desta maquinaria é importante para compreender o enovelamento proteico no meio intracelular. As hsp70 são proteínas menores, que se ligam a sequências hidrofóbicas expostas e mantêm a cadeia peptídica desenovelada até que ela possa assumir a estrutura tridimensional correta. Esta chaperona tem duas tarefas importantes: ajudar o enovelamento e impedir que várias proteínas malformadas, com sequências hidrofóbicas expostas, formem agregados que, além de inúteis, podem ser muito nocivos. Foram reportadas hsp70 mitocondriais nas espécies *L. chagasi* e *L. infantum* (Campos *et al*. 2008).

A maior protease de superfície, a GP63, é a mais abundante glicoproteína em *Leishmania spp*. Sua aplicação em clínica tem crescido substancialmente e novas drogas antiparasitárias vêm sendo desenvolvidas, incluindo antígenos, os quais poderão ser testados como vacina para a leishmaniose e também para a melhoria da terapêutica e do diagnóstico dessas doenças. Os parasitas que podem estabelecer e manter uma infecção em um hospedeiro mamífero utilizam vários métodos para evitar os efeitos prejudiciais da resposta imune do hospedeiro. O gênero *Leishmania* usa a GP63 para invasão celular e como um meio de sobreviver a mecanismos imunes do hospedeiro. As funções da GP63 em *Leishmania* como uma metaloprotease zinco são

desempenhadas na superfície celular e proporcionam resistência à lise mediada pelo complemento. Homólogos do RNAm para a metaloprotease zinco já foram identificados na espécie *T. cruzi*. As sequências das proteínas GP63 preditas em *T. cruzi* mostram similaridades de 37% e 43% com as proteínas GP63 de *L. guyanensis* e *T. brucei*, respectivamente. A identificação de múltiplos genes de GP63 em *T. cruzi* fornece uma oportunidade para estudar os possíveis papéis da proteína na sobrevivência e infectividade deste parasita (Mauricio *et al*. 2007).

A enzima tripanotiona redutase foi descoberta em 1985 e atualmente é considerada um alvo molecular validado para o planejamento de inibidores a serem utilizados no tratamento da doença de Chagas. Esta enzima é dependente de NADPH e catalisa a redução da tripanotiona disulfeto [T(S)2] em tripanotiona ditiol [T(SH)2], desencadeando, assim, uma cascata de eventos responsáveis pela neutralização de espécies reativas de oxigênio. Desta forma, a tripanotiona redutase mantém um ambiente redutor no interior do parasita, protegendo-o contra o estresse oxidativo. Esta enzima tripanotiona redutase é membro da família das enzimas dissulfeto oxidoredutase e encontra-se apenas em protozoários parasitas dos gêneros *Trypanosoma* e *Leishmania* (Singh *et al*. 2008). Além de exercer uma função fundamental para o parasita, outra característica que faz com que a tripanotiona redutase seja um alvo interessante para o planejamento de fármacos antiparasitários é sua significativa diferença estrutural em relação à glutationa redutase, enzima com função correspondente no homem. As principais diferenças entre a tripanotiona redutase e a glutationa redutase estão relacionadas ao tamanho, carga e distribuição dos bolsões hidrofílicos/hidrofóbicos. Estas diferenças fazem com que o sítio ativo da tripanotiona redutase, em comparação com o da glutationa redutase, seja capaz de acomodar facilmente substratos apresentando grupos mais volumosos. Além disso, a tripanotiona redutase apresenta resíduos carregados negativamente e regiões hidrofóbicas em seu sítio ativo, os quais são capazes de interagir, via ligações eletrostáticas e de van der Waals, com seu substrato. A glutationa redutase , por sua vez, apresenta resíduos de arginina, carregados positivamente, que interagem por ligação eletrostática com seu substrato. Estas diferenças podem ser aproveitadas no planejamento de inibidores seletivos para a tripanotiona redutase (Oliveira *et al*. 2008).

Têm sido feitas várias análises sobre o papel dos transportadores ABC (*ATP-Binding Cassette*) na resistência às drogas. Estas proteínas de membrana medeiam o transporte dependente de ATP de uma ampla variedade de drogas quimioterápicas aos alvos dentro dos

parasitas. Os transportadores da família ABC são conhecidos por estabelecerem a base da resistência a múltiplas drogas em células de câncer em mamíferos e leveduras patogênicas, fungos e bactérias. Estas proteínas constituem uma das maiores famílias de proteínas de membrana encontradas nos procariotos e eucariotos. Alguns dos mais conhecidos transportadores ABC associados à resistência clínica em humanos são a glicoproteína P (Pgp), codificada pelo gene MDR1, e alguns membros do subgrupo das Proteínas Associadas à Resistência a Multidrogas, codificadas pelo gene MRP1-8. Estas observações foram determinantes para as principais pesquisas sobre transportadores ABC tipo MDR e MRP em protozoários parasitas (Klokouzas *et al*. 2003; Mukherjee *et al*. 2007; Ouellette *et al*. 2001; Sauvage *et al*. 2009).

1.2 Filogenômica baseada na reconstrução da árvore de espécies de protozoários

Diversas dificuldades são encontradas na busca de uma definição taxonômica consenso dos diversos domínios. Em particular, os protozoários estão vagamente caracterizados. Decidir se uma espécie pertence aos protozoários ou não depende da morfologia e das propriedades biológicas e também, em parte, do fato de não pertencerem a outro reino (Turmel *et al*. 2002) (Cavalier-Smith 2005; Levine *et al*. 1980). Os protozoários não são um grupo filogenético coerente como outros reinos ou reinos candidatos (exemplo: stramenopilas). Assim, para determinar se uma espécie pertence aos protozoários, não basta olhar só para uma característica (como os cabelos tubulares flagelares de Stramenopila) e decidir. Vários critérios têm de ser cumpridos.

Os protozoários estão filogeneticamente ligados a outros grupos eucarióticos, e diversos grupos eucarióticos são provavelmente derivados de protozoários (Vickerman 1976). Um exemplo disso é a relação estabelecida por similaridade molecular entre stramenopilas e protozoários alveolatas (Cavalier-Smith 1999).

Outro exemplo é a relação entre plastídeos residuais das espécies apicomplexas dos gêneros *Plasmodium* e *Toxoplasma* com plastídeos dinoflagelados e/ou plastídeos de cromistas stramenopilas (Cavalier-Smith 1999). Os dados moleculares relacionam os protozoários coanoflagelados com esponjas, e, por conseguinte, com o reino Animalia. Curiosamente, o sequenciamento molecular também suporta uma relação de coanoflagelados com o reino Fungi (Vickerman 1976). Sequências RNA ribossomais SSU apontam para um agrupamento entre

Biliphyta (glaucofitas, algas vermelhas) e criptomonades. Além disso, Nakayama *et al*., em 1998, forneceu provas moleculares mostrando que protozoários flagelados primitivos ocorreram na ascendência das clorofitas (algas verdes).

Das dez doenças especificadas como prioridades de investigação pelo Programa Especial da Organização Mundial da Saúde para Pesquisa e Treinamento em Doenças Tropicais (http://www.who.int/tdr), quatro são causadas por protozoários parasitas (malária, leishmaniose, doença de Chagas, tripanossomíase africana).

Estas doenças e outras menos perigosas (exemplos: amebíase e tricomoníase) estão tendo um aumento alarmante de casos de refratividade ao tratamento principal. O fracasso do tratamento tem, potencialmente, uma origem multifatorial, com uma das principais preocupações: a resistência a drogas (Arango *et al*. 2008; Arevalo *et al*. 2007; Bansal *et al*. 2004; Burri & Keiser 2001; Sobel *et al*. 1999).

Nossa compreensão da posição filogenética dos protozoários em eucariotos, bem como suas relações, é baseada principalmente na análise do RNA ribossômico (Wainright *et al*. 1993). Inicialmente, este processo parecia relativamente simples: as árvores geradas a partir de um único gene, geralmente uma pequena subunidade do gene do RNA ribossômico (Sogin 1991), pareciam fornecer uma estrutura de base para a topologia de eucariotos, embora muitos ramos da árvore permanecessem controversos. A década de 1990 foi um período de desconstrução desta teoria, pois diversas árvores de genes que codificam proteínas revelaram discrepâncias graves (Keeling *et al*. 1996; Simpson *et al*. 2002). Atualmente, a árvore aceita - a árvore de eucariotas - assemelha-se à apresentada por Keeling em 2005. A árvore de eucariotas é composta de vários tipos de dados, incluindo filogenias moleculares e outras características moleculares, bem como evidências morfológicas e bioquímicas. Cinco "supergrupos" mostraram, cada um, estar composto por uma diversidade de eucariotas, a maioria das quais composta principalmente por protistas e algas. Atualmente, a árvore aceita - a árvore de eucariotos - assemelha-se à apresentada por Keeling em 2005 e é composta de vários tipos de dados, incluindo filogenias e outras características moleculares, bem como evidências morfológicas e bioquímicas. Cinco "supergrupos" mostraram, cada um, estar compostos por uma diversidade de eucariotos, a maioria dos quais composta principalmente por protistas e algas.

Chaudhary, em 2005, fez uma analise filogenética que definiu cinco supergrupos de eucariotos:

(i) a linhagem de plantas e algas vermelho/verde,

(ii) um clado formado por animais, fungos, *slime molds* e Amoebozoa e três grupos que são inteiramente de protozoários,

(iii) Chromalveolatas

(iv) Excavatas e

(v) Rhizaria (Keeling *et al*. 2005).

Os protozoários incluídos neste estudo se apresentam distribuídos em três grupos:

- grupo 1: **phylum** parasita Apicomplexa (*Theileria*, *Babesia*, *Plasmodium*, *Toxoplasma*, *Neospora*, *Eimeria* e *Cryptosporidium*),
  ciliados (*Tetrahymena*, *Paramecium* e *Oxytricha*),
  diatomáceas e muitos táxons para os quais genomas não completos estão disponíveis (*Phytophtora*) (Stechmann *et al*. 2003a; Stechmann *et al*. 2003b).
- grupo 2: **excavatas**, formado pelos parasitas Kinetoplastida (tripanossomas africanos, tripanossomas sulamericamos e *Leishmania*) e outras linhagens (*Giardia*, *Spironucleus*, *Trichomonas* e *Nagleria*) (Simpson *et al*. 2002).
- grupo 3: **eucariotas Unikonta** (Amoebozoa e Opisthokonta), que continham espécies incluídas no presente estudo (*Acanthamoeba*, *Entamoeba* e *Dictyostelium*) (Cavalier-Smith 2002).

1.3 Filogenômica dos EGM em protozoários

1.3.1 Genômica comparativa

1.3.1.1 Comparação por similaridade

Sequências homólogas são aquelas que compartilham o mesmo ancestral e que, provavelmente, têm a mesma função. O processo de anotação de sequências é auxiliado pela comparação entre elas e bancos de dados como GENBANK (Benson *et al*. 2004), TREMBL (Boeckmann *et al*. 2003) e SWISS-PROT (Bairoch *et al*. 2004), o que tornou imprescindível o estabelecimento de métodos de genômica comparativa para entender e esclarecer o processo evolutivo e as relações filogenéticas das diversas espécies. As informações genéticas geradas pelos projetos genomas tornam necessário o desenvolvimento de ferramentas computacionais com o objetivo de auxiliar na interpretação dos dados genômicos.

A comparação por similaridade tem sido o método mais utilizado na detecção de homologias. Estes algoritmos de comparação baseiam-se principalmente em técnicas de programação dinâmica, como descrito por Smith e Waterman em 1981. Algumas ferramentas populares, como os pacotes BLAST (Altschul *et al*. 1990) e FASTA (Pearson 1990) são implementações desses algoritmos.

Porém, a maioria dos pacotes, tais como o HMMER (Eddy 1998), o SAM (Hughey *et al*. 1996) e o THMM (Qian *et al*. 2004), criam perfis (*profiles*) *Hidden Markov models* a partir de um conjunto de sequências homólogas e avaliam o quanto outras sequências se adaptam ao perfil. Particularmente, os métodos baseados no algoritmo *Hidden Markov models* (HMM) utilizam um tipo especial de modelo probabilístico, denominado perfil *Hidden Markov models* (pMM) e que tem demonstrado uma alta eficiência na detecção de homologias distantes (Gough *et al*. 2001; Madera *et al*. 2002; Park *et al*. 1998; Wistrand *et al*. 2005).

1.3.1.2 Alinhamento múltiplo

O alinhamento múltiplo de sequências é considerado o passo mais importante em muitas aplicações na bioinformática. A eleição de um algoritmo de programa de alinhamento múltiplo, com ótimo desempenho computacional, produz efeitos benéficos para o melhor alinhamento das sequências biológicas (Park *et al*. 1998; Wallace *et al*. 2005).

O alinhamento múltiplo de tipo progressivo (Taylor 1988) é o método heurístico mais usado para alinhar um grande número de sequências e é construído progressivamente alinhando pares de sequências seguido de pares de alinhamento. Esta técnica é usada em muitos programas diferentes de alinhamento múltiplo, como o MULTIALIGN (Barton *et al*. 1987), o CLUSTALW (Thompson *et al*. 1994) e o T-COFFEE (Notredame *et al*. 2000).

Um dos problemas com os alinhamentos progressivos é que não existe um mecanismo que faça uma posterior correção dos erros introduzidos no início do processo de alinhamento. No entanto, para corrigir este problema, Barton e Sternberg, em 1987, no programa MULTIALIGN, recorreram ao processo de interação. Posteriormente o programa PRRP (Gotoh 1996) usou uma estratégia de interação muito mais sofisticada. Entre os programas de alinhamento múltiplo mais usados estão o PROBCONS (Do *et al*. 2005) e o MUSCLE (Edgar 2004) que refinam o alinhamento múltiplo final usando a interação. O PROBCONS implementa um "algoritmo de

particionamento randômico", enquanto o MUSCLE implementa um "algoritmo de particionamento baseado em uma árvore" para a interação.

Atualmente, o programa MAFFT é um dos melhores programas de alinhamento múltiplo de sequências e tem sido usado em recentes publicações devido à eficiência e rapidez na reconstrução dos alinhamentos. O tempo que leva a execução do MAFFT na Unidade Central de Processamento (CPU) é drasticamente reduzido quando comparado ao de outros métodos existentes. O MAFFT inclui duas novas técnicas: (i) regiões homólogas são rapidamente identificadas pelo método *fast Fourier transform* (FFT), em que uma sequência de aminoácidos é convertida a uma sequência composta de volume e valores de polaridade de cada resíduo do aminoácido e (ii) um sistema simplificado de contabilização que reduz o tempo de processamento, aumentando a precisão dos alinhamentos tanto para sequências com grandes inserções, como para sequências distantemente relacionadas, mas com comprimentos semelhantes. Dois algoritmos heurísticos diferentes - o método progressivo (FFT-NS-2) e o método de refinamento interativo (FFT-NS-I) - foram implementados no MAFFT. Compararam-se os desempenhos de FFT-NS-2 e FFT-NS-I com outros métodos de simulações em computador e testes de *benchmark*. Resultado: o tempo de processamento de FFT-NS-2 foi drasticamente reduzido quando comparado com o do CLUSTALW.

1.3.2 Filogenia e filogenômica

A diversidade biológica, ou seja, as diferenças entre os grupos de organismos são entendidas hoje como o resultado do processo de evolução. Os seres vivos não são entidades estáticas. Eles se transformam ao longo das gerações sob influência do meio. A Sistemática é um ramo da biologia que tem como objeto a descrição dessa diversidade. A partir das comparações entre as características observadas nos organismos atuais e nos fósseis, os pesquisadores procuram estabelecer relações de parentesco e ancestralidade.

O método filogenético consiste basicamente na separação de organismos em grupos que compartilham propriedades específicas. A classificação assim obtida pode ser representada em uma estrutura de árvore conhecida como árvore filogenética (filogenia). Os nós de uma árvore filogenética correspondem a entidades biológicas (espécies, famílias ou fragmentos de DNA) e as arestas, as mudanças de propriedades entre estes nós. Nas árvores filogenéticas, as transições ocorridas nas arestas agrupam as subárvores cujos nós compartilham as propriedades derivadas.

A construção da filogenia, no entanto, pode ser feita computacionalmente, se forem fixados alguns critérios que definam como deve ser uma árvore que represente bem os dados. Os problemas presentes neste aspecto computacional são de interesse direto para cientistas da computação e têm sido exaustivamente estudados. Existem hoje diversos algoritmos e programas que fornecem bons resultados e são amplamente utilizados por bioinformatas.

A filogenia ajuda a inferir a história evolutiva das espécies hoje existentes e verifica os relacionamentos entre estas espécies, a fim de determinar possíveis ancestrais comuns entre elas. Uma árvore filogenética é uma árvore onde as folhas representam os organismos e os nós internos, seus supostos ancestrais. As arestas da árvore indicam as relações evolutivas. Podem ser construídas árvores filogenéticas para populações, espécies, gêneros ou outros grupos de indivíduos, inclusive utilizando dados morfológicos descritivos ou sequências de proteínas e de ácidos nucléicos. Atualmente podem ser usadas sequências de dados disponíveis em diversos bancos de dados gerados por muitos projetos genomas concluídos e outros ainda em andamento.

1.3.2.1 Filogenômica

A filogenômica tem sido definida essencialmente como a interseção entre a evolução e a genômica (Delsuc *et al*. 2005; Graham *et al*. 2004; O'Brien *et al*. 1999; Pellegrini *et al*. 1999; Sicheritz-Ponten *et al*. 2001; Sjolander 2004). O termo tem sido usado para se referir a análises relacionadas a dados genômicos e a reconstruções evolutivas, especialmente filogenéticas. Outra definição (Brown *et al*. 2006; Conte *et al*. 2008; Daubin *et al*. 2002; Dopazo *et al*. 2004; Glanville *et al*. 2007; Krishnamurthy *et al*. 2005; Venter *et al*. 2001; Zmasek *et al*. 2002; Zmasek *et al*. 2001) é a reconstrução da vida evolutiva dos organismos com base nas análises dos seus genomas. No entanto, a inferência filogenômica também se refere à análise detalhada de famílias de proteínas individuais (Citerne *et al*. 2003; Gadelle *et al*. 2003; Jeffroy *et al*. 2006; Marcotte *et al*. 2000; Vienne *et al*. 2003).

Entre as aplicações atribuídas à filogenômica estão: (i) a predição da função de um gene baseada na história evolutiva, representada em uma árvore filogenética, idéia original de Jonathan Eisen (Eisen 1998; Eisen *et al*. 2003; Eisen *et al*. 1999; Eisen *et al*. 2002), (ii) a reconstrução de uma árvore de espécies combinando informações de vários genes ou genomas inteiros e (iii) a integração de análises genômicas para a reconstrução evolutiva (Wu *et al*. 2008).

Reconstruir as relações filogenéticas entre todos os organismos vivos - visando entender os processos evolutivos na vida das espécies - é um dos desafios fundamentais da biologia. Tem sido mencionado, que "nada na biologia faz sentido exceto à luz da evolução" e que "nada na biologia evolutiva faz sentido exceto à luz de uma filogenia" (Savage 1977) (Dobzhansky 1973).

Tipicamente, a filogenia molecular envolve três etapas: (i) a recuperação de sequências homólogas, (ii) a criação de alinhamentos múltiplos de sequências e (iii) a construção de árvores filogenéticas.

Algoritmos de agrupamento de vizinhos (AV) (Saitou *et al*. 1987), máxima parcimônia (MP) (Felsenstein 1996; Swofford 2002; Wilgenbusch *et al*. 2003; Yang 1994), máxima verossimilhança (MV) (Felsenstein 1981; Felsenstein 1996) e inferência bayesiana (IB) (Holder *et al*. 2003) têm sido desenvolvidos para calcular e reconstruir a árvore na qual as relações de similaridade/herança entre as sequências sejam mais bem refletidas (Felsenstein 2003; Thornton *et al*. 2000).

Por outro lado, quando se tenta inferir filogenia, existem problemas potenciais ou incongruências que poderiam ser o resultado de: (i) violações das hipóteses de ortologia geradas por mecanismos como duplicação de genes, transferência horizontal de genes ou linhagem das espécies escolhidas, (ii) erros estocásticos em relação ao comprimento dos genes e (iii) erros sistemáticos na reconstrução da árvore devido a artefatos gerados pela presença de um sinal não filogenético nos dados (Daubin *et al*. 2003; Doolittle 1999; Dutilh *et al*. 2004; Jeffroy *et al*. 2006).

A filogenômica tem sido intensamente utilizada na "Era Pós Genômica". A abordagem filogenômica faz uso de um grande número de genes descobertos por projetos genoma e/ou pelo sequenciamento dos EST (*expressed sequence tags*). Os maiores alinhamentos pré-filogenômicos consistiram em cerca de 10 genes. Hoje, um típico alinhamento filogenômico compreende entre 50 e várias centenas de genes (Telford 2007).

Embora, os genes RNA ribossomais e outros genes únicos tenham sido extremamente valiosos para estudos filogenéticos, a filogenia gene-único tem suas limitações (Teichmann *et al*. 1999; Wu & Eisen 2008). Portanto, a abordagem filogenômica, corroborada pela utilização dos marcadores filogenéticos mais representativos, permitirá, em tese, uma inferência mais fiável e representativa para a árvore da vida.

A filogenômica envolve o uso de genomas inteiros para inferir filogenia de uma árvore de espécies e tornou-se o padrão para reconstruir filogenias de espécies confiáveis (Ciccarelli *et al*. 2006; Daubin *et al*. 2002). No entanto, a recente árvore do superalinhamento (Ciccarelli *et al*. 2006) tem recebido algumas críticas como sendo uma "árvore de um por cento" do genoma (Dagan *et al*. 2006), considerando que as árvores filogenéticas de um único gene mostraram conflito (Teichmann & Mitchison 1999) devido a uma variedade de causas. Apesar disso, as árvores filogenômicas têm mantido a promessa de minimizar as anomalias pela força dos dados em larga escala genômica, ao fato de serem baseadas na quantidade máxima de informação genética. Uma árvore filogenômica teria de ser o melhor reflexo da história evolutiva da espécie (Doolittle 1999; Ge *et al*. 2005).

Atualmente, tem-se a opção de concatenar sequências de genes múltiplos com mais sinais filogenéticos para construir árvores no nível genômico, como as "árvores do genoma" - também chamadas "árvores supermatriz". Estas árvores são menos suscetíveis a erros estocásticos (Daubin *et al*. 2003; Doolittle 1999; Dutilh *et al*. 2004; Jeffroy *et al*. 2006) do que as árvores construídas a partir de um único gene.

A outra opção é a construção da superárvore que envolve a concatenação de um conjunto de árvores (Ciccarelli *et al*. 2006; Creevey *et al*. 2005; Huerta-Cepas *et al*. 2007). No entanto, existem diferenças fundamentais entre as formas em que abordagens filogenômicas integram as informações filogenéticas. Dutilh *et al*., em 2007, comparou de forma sistemática metodologias alternativas tais como: o conteúdo de genes, os superalinhamentos, as superdistâncias (para construir uma supermatriz) e as abordagens de superárvores utilizando vários algoritmos e métodos de construção de árvore no domínio Fungi (clado eucariótico com o maior número de genomas sequenciados). Segundo os mesmos autores, as árvores filogenômicas reproduziram muitos dos clados de acordo com as atuais visões taxonômicas. Os superalinhamentos (supermatrizes) e superárvores reproduziram a melhor filogenia de fungos - métodos fenotípicos tradicionais (Guarro *et al*. 1999), análises filogenéticas moleculares baseadas em RNAr (Fell *et al*. 2000; Scorzetti *et al*. 2002; Tehler *et al*. 2003) ou pequenos números de outras proteínas (Diezmann *et al*. 2004; James *et al*. 2006; Kouvelis *et al*. 2004; Kurtzman 2003), assim como também alguns estudos em grande escala (Jeffroy *et al*. 2006; Kuramae *et al*. 2006; Robbertse *et al*. 2006; Rokas *et al*. 2003; Thomarat *et al*. 2004) -, mas não foram garantia de uma árvore filogenômica bem sucedida (Dutilh *et al*. 2007).

1.3.2.2 Algoritmos filogenéticos

Desde 1980, quando a primeira versão do programa PHYLIP (*PHYLogeny Inference Package*) (Felsenstein 2005) foi introduzida por Felsenstein, um elevado número de programas filogenéticos vem sendo desenvolvido.

Atualmente, os programas PHYLIP (Felsenstein 2005), PAUP (Swofford 2002; Wilgenbusch & Swofford 2003), MEGA (Kumar *et al*. 2008), PHYML (Guindon *et al*. 2003; Guindon *et al*. 2005), PAML (Yang 2007) e MRBAYES (Yang 2007) estão entre os mais bem conhecidos e encontram-se disponíveis para milhares de usuários ao redor do mundo (Tarraga *et al*. 2007). As novas versões paralelizadas e de velocidade otimizada de alguns programas - MRBAYES, RAXML – tornaram-se a melhor opção para a análise de grandes quantidades de dados (Fuellen 2008).

Os novos programas desenvolvidos e executados através de algoritmos sofisticados tornam as análises filogenéticas mais promissoras. Entre estes programas encontram-se: GARLI (Zwickl 2006), RAXML (Stamatakis *et al*. 2005), MAFFT (Keane *et al*. 2006), AWTY (Nylander *et al*. 2008), BEAST (Drummond *et al*. 2007), LAMARC (Kuhner 2006), MIGRATE-N (Carstens *et al*. 2005) e HYPHY (Pond *et al*. 2005).

Atualmente, tem-se a opção de executar os programas de forma local em um servidor próprio - através da linha de comando de um terminal - ou por meio de um servidor externo, usando uma interface *web*. O servidor *web* disponibiliza a execução de programas de bioinformática que vão desde:

- o cômputo de um alinhamento múltiplo usando o CLUSTALW (http://www.ebi.ac.uk/clustalw/) ou o MAFFT (http://www.imtech.res.in/raghava/mafft/),

- a construção de modelos evolutivos usando o MODELGENERATOR (http://distributed.cs.nuim.ie/multiphyl.php) e

- a construção de árvores filogenéticas usando o PHYML (http://atgc.lirmm.fr/phyml/).

1.3.2.3 *Pipelines* automatizados para a homologia, a inferência de árvores filogenéticas e a anotação funcional

A filogenômica examina cuidadosamente as anotações funcionais das sequências em grande escala. Assim, uma vez que estabelece quais sequências intervirão em um estudo, utiliza-as na construção de árvores filogenéticas (Eisen 1998; Eisen & Wu 2002; Rannala *et al*. 1998).

Geralmente os *pipelines* filogenômicos tentam automatizar quatro tarefas: (i) coletar sequências informativas, (ii) alinhá-las, (iii) gerar uma árvore ou um conjunto de árvores e (iv) analisar as informações evolutivas e filogenéticas.

A análise da crescente quantidade de dados biológicos e biomédicos pode ser comparada dentro das espécies e entre elas. Por exemplo: uma análise integrativa dos dados de projetos de sequenciamento de genomas de diversas espécies infere a evolução dos genomas e identifica as partes conservadas e as polimórficas. Os dados biológicos são comparáveis se existe uma origem comum - homologia – como é o caso dos membros de uma família de genes originários da duplicação de um gene ancestral. Se a família de genes tem parentes em outras espécies, podemos supor que o gene ancestral estava presente na espécie ancestral a partir da qual todas as outras espécies evoluíram.

Inferir a origem comum ancestral pode explicar como certas sequências biológicas vêm se desenvolvendo e pode prever quais sequências pertencem a uma determinada espécie e qual é a sua função. Essa transferência de informação de sequências em uma espécie para sequências homólogas de outras espécies é baseada no princípio "é meu parente mais próximo e comporta-se como minha função", muitas vezes referida como "culpa por associação".

Para permitir a transferência de conhecimentos em grande escala, vários *pipelines* filogenômicos automatizados têm sido desenvolvidos nos últimos anos. A revisão descrita por Fuellen em 2008 - baseada na descrição e comparação de sete *pipelines* filogenômicos - demonstrou que a homologia e as análises filogenéticas feitas e automatizadas em grande escala possibilitam uma melhor compreensão sobre a função, distribuição e relação das sequências em determinadas espécies.

Outra ferramenta filogenômica automatizada é a pyphy. Formada por esquemas que representam a estrutura da árvore e que estão relacionados entre si por meio de "conexões filogenéticas", esta ferramenta encontra-se disponível apenas para o domínio Archaea (Sicheritz-Ponten & Andersson 2001). Em seguida à pyphy, surgiu o programa RIO (*resampled inference of orthologs*) fornecendo uma estimativa para a ortologia e a paralogia (Zmasek & Eddy 2002). Uma outra ferramenta automatizada – a PipeAlign - recupera dados de sequências relacionadas de bancos de dados e gera alinhamentos múltiplos (Plewniak *et al*. 2003).

O programa PhyloGenie complementou o trabalho da ferramenta pyphy através da automatização da análise e geração dos filomas, mantendo um banco de dados de árvores

(Frickey *et al*. 2004). Um outro programa – o Figenix - calcula as árvores usando os algoritmos AV, MP e MV e fornece uma automatização inteligente na anotação genômica (Gouret *et al*. 2005).

O pacote BLAST baseado no procedimento RiPE (*retrieval-induced phylogeny estimation*) (Fuellen *et al*. 2005) automatiza as análises filogenômicas e foi utilizado para analisar a evolução das proteínas ABC (Fuellen *et al*. 2005; Spitzer 2006) e FinGER (Stolle *et al*. 2005), das polimerases *RNA DNA-directed* (Klenk *et al*. 2004), das proteínas S100 e das tirosina quinases (Spitzer 2006).

O Grupo de Filogenômica de Berkeley (http://phylogenomics.berkeley.edu) fornece um conjunto de servidores *web* para a execução dos seguintes principais programas e algoritmos filogenômicos (Glanville *et al*. 2007):

- a enciclopédia filogenômica PhyloFacts, que contém uma classificação pré-computada de sequências de famílias e subfamílias de genes (Krishnamurthy *et al*. 2006),

- o algoritmo de clusterização FlowerPower, projetado para a identificar homólogos e realizar a análise filogenômica estrutural (Krishnamurthy *et al*. 2007),

- os programas de alinhamento múltiplo de sequências MUSCLE e SATCHMO (Edgar *et al*. 2003a; Edgar *et al*. 2003b),

- o algoritmo SCI-PHY, para a identificação de subfamílias de genes (Krishnamurthy *et al*. 2007) e

- o servidor *web* PhyloBuilder que fornece um *pipeline* filogenômico integrado. A execução começa com o fornecimento da sequência de proteína pelo usuário, seguido da identificação dos homólogos, construção do alinhamento múltiplo, construção da árvore filogenética, identificação das subfamílias e predição da estrutura.

Outro servidor *web* é o Phylemon, que reúne um conjunto de diferentes ferramentas filogenéticas e evolutivas, oferecendo um ambiente integrado que permite a concatenação direta de análises bioinformáticas multipasso (Tarraga *et al*. 2007).

1.3.2.4 O sistema ARPA na análise filogenômica

Sabe-se que o uso de abordagens integradas - como o uso de um único *pipeline (*por exemplo, o GARSA) (Davila *et al*. 2005) - ajuda na análise e na anotação de genes. Esta metodologia também é válida na identificação e nas análises de reconstrução filogenética e

evolutiva dos EGM, foco deste estudo. Um *pipeline* usado para a anotação dos EGM na espécie *D. melanogaster* (Quesneville *et al*. 2005) integra os resultados dos seguintes métodos/programas baseados em homologia: o RepeatMasker, o BLASTER, o TBLASTX, o *all-by-all* BLASTN, o RECON e o TE-HMM. Outro sistema para a caracterização de *retroids* no genoma humano é o *Genome Parsing Suite*, (McClure *et al*. 2005).

Novas abordagens filogenéticas são indispensáveis para genes pouco conservados, sendo necessária uma metodologia que vise usar de maneira apropriada vários algoritmos filogenéticos e evolutivos com seus respectivos parâmetros. A filogenia de genes codificados pelos EGM em protozoários pode ser difícil de ser inferida, pois estes EGM apresentam baixos níveis de conservação nas sequências. A isto se somam os poucos genomas inteiramente sequenciados, o que torna este estudo difícil, porém original.

Entre os algoritmos filogenéticos mais usados estão: o agrupamento de vizinhos (Saitou & Nei 1987), a máxima parcimônia (Felsenstein 1996; Swofford 2002; Wilgenbusch & Swofford 2003; Yang 1994), a máxima verossimilhança (Felsenstein 1981; Felsenstein 1996) e a inferência Bayesiana (Holder & Lewis 2003), além de outros derivados destes algoritmos.

A filogenômica dos tripanossomatídeos e protozoários foi executada usando o sistema ARPA. Este sistema é composto por 18 programas de bioinformática - oito de filogenia - estruturados de forma que permite uma execução automática, e pode ser executado via interface *web* ou localmente via linha de comando através de um terminal. O sistema ARPA assim como sua interface estão hospedados no Instituto Oswaldo Cruz - FIOCRUZ e podem ser acessados no endereço http://arpa.biowebdb.org.

1.3.3 Elementos genéticos móveis (EGM)

Os EGM são sequências de DNA com a habilidade de se replicarem dentro do genoma do hospedeiro (Craig *et al*. 2002). Usualmente se comportam como parasitas intragenômicos (consistentes com o modelo de "DNA egoísta") (Doolittle *et al*. 1980; Frost *et al*. 2005; McClintock 1941a; McClintock 1938; McClintock 1941b; Orgel *et al*. 1980).

Os EGM incluem duas classes:

(i) os Elementos Classe I ou retrotransposons, os quais usam uma molécula de RNA como intermediária no processo de retrotransposição. A retrotransposição é mediada pela enzima chave transcriptase reversa e conhecida comumente como um mecanismo de "copia e pega" e

(ii) os Elementos Classe II ou transposons, os quais usam uma molécula de DNA como intermediária no processo de transposição. A transposição é mediada pela enzima chave transposase e conhecida comumente como o mecanismo de "corta e pega" (Finnegan 1989; Holmes 2002).

O termo transposição tem sido muito usado em genética para descrever a transferência de segmentos cromossômicos nos rearranjos estruturais. Estes segmentos cromossômicos que estariam sendo transpostos seriam um gene, um pequeno número de genes ligados ou um fragmento do tamanho de um gene (figura 1.1). Acredita-se que a origem da vida provém do "mundo do RNA", seguido por uma transcrição reversa para o DNA. Neste mundo, os EGM poderiam ter sido participantes ancestrais ativos (Brosius *et al*. 1995). Estudos publicados referenciam alguns dos principais genes codificados pelos EGM: (i) transcriptase reversa (Eickbush 1994; Xiong *et al*. 1993; Xiong *et al*. 1990; Xiong *et al*. 1988), (ii) telomerase (Adams *et al*. 2000; Anzai *et al*. 2001; Biessmann *et al*. 1990; Levis *et al*. 1993; Noutoshi *et al*. 1998; Pich *et al*. 1998; Takahashi *et al*. 1997; Zhu *et al*. 1999), (iii) ribonuclease H (Chapados *et al*. 2001; Kashiwagi *et al*. 1996; Malik *et al*. 2001; Misra *et al*. 2005; Oda *et al*. 1993), (iv) proteína env, (v) proteína gag e (vi) proteína pol (Covey 1986; Fawcett *et al*. 1986; Xiong & Eickbush 1988).

Os EGM são partes integrais dos genomas eucarióticos (Novikova *et al*. 2009) e devido a capacidade de transposição e o grande número de cópias, os EGM deveriam exercer um papel importante na organização, função, regulação assim como também na evolução dos genes (Bennetzen 2000; Kidwell *et al*. 2001). A presença e atividade dos EGM proporcionam plasticidade ao genoma, pois eles participam na integração dos pseudogenes e o *exon shuffling*.

Os EGM estão presentes na maioria dos genomas. Em mamíferos, eles e seus remanescentes reconhecíveis fazem parte de aproximadamente metade do genoma (*Genome Sequencing Consortium of Mouse (2002) and Human (2001)*), e em algumas plantas constituem mais de 90% do genoma (SanMiguel *et al*. 1996).

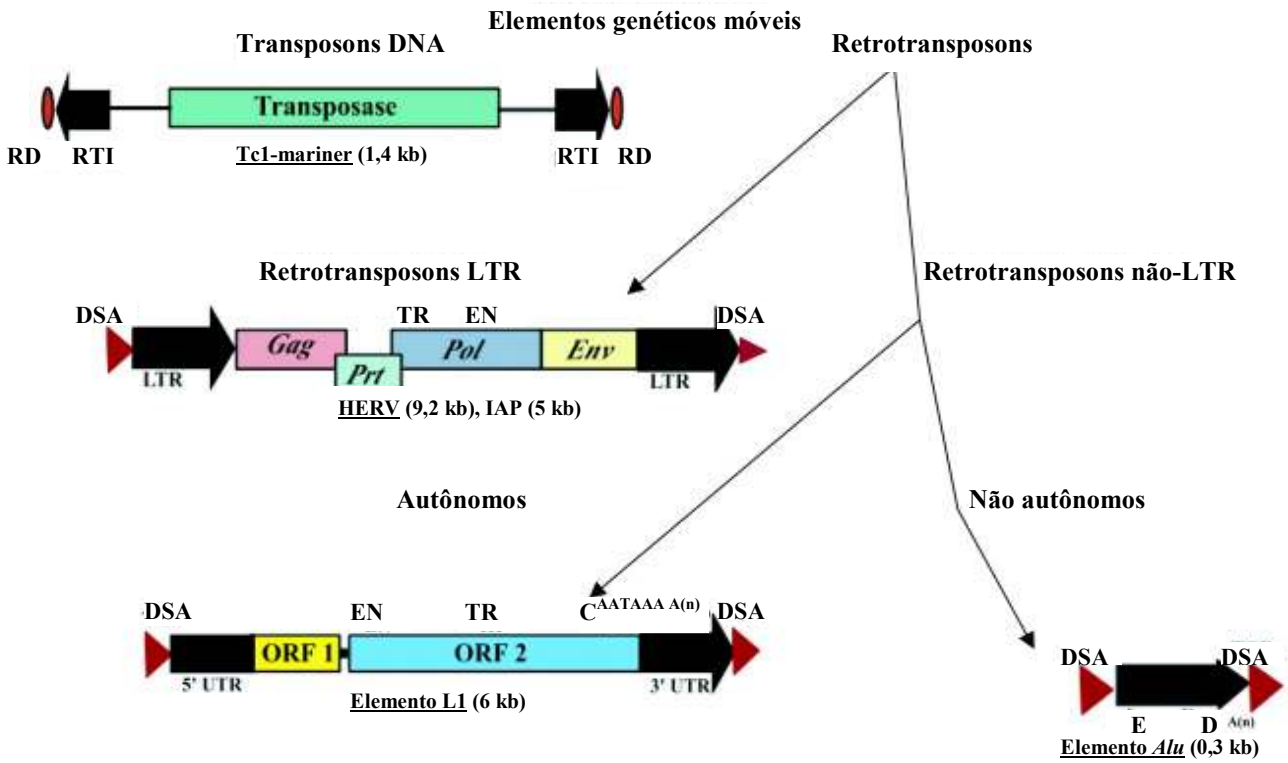


Figura. 1.1 As classes e a estrutura organizacional dos EGM

Os transposons DNA codificam a enzima transposase e apresentam repetições diretas (RD) e repetições terminais invertidas (RTI) nas extremidades, tendo como exemplo o elemento Tcl-mariner. Os retrotransposons apresentam duas classes maiores: Os LTR e os não-LTR. Os retrotransposons LTR codificam um conjunto de enzimas - proteína gag, protease, proteína pol, proteína env, transcriptase reversa e endonuclease - e possuem nas extremidades duplicações de sítio-alvo (DSA) e repetições terminais longas - LTR (*long terminal repeats*) - como pode-se notar nos elementos HERV e IAP. Os retrotransposons não-LTR podem ser classificados em dois grupos: autônomos e não autônomos, dependendo da capacidade de codificação das enzimas. Os autônomos apresentam duas ORF (*open reading frame*) que podem estar relacionadas às enzimas endonuclease e transcriptase reversa. Além disso, possuem uma cauda poli-A e as DSA (elemento L1). No entanto, os elementos não autônomos não codificam enzima alguma e não apresentam ORF. Por outro lado, pode-se notar a presença das DSA no elemento Alu. Fonte Kazazian Jr. HH (2004).

Uma característica dos EGM é a alta probabilidade de eliminação da população por perda estocástica (extinção vertical). Uma das estratégias que possibilitam a persistência dos EGM intra ou intergenomas é a habilidade de "pular" o limite de classificação de espécies e se propagar entre elas (transferência horizontal) (Lohe *et al*. 1997). Embora exista evidência de transferência

horizontal dos EGM entre muitas espécies diferentes, ainda não há uma análise sistemática que avalie a prevalência deste processo nas famílias dos EGM em genomas de organismos modelos (Sanchez-Gracia *et al*. 2005).

Uma busca por genes da transcriptase reversa resultou na identificação de novos retrotransposons no genoma da *Drosophila melanogaster* (Berezikov *et al*. 2000), o que sugere que buscas sistemáticas dos genes envolvidos na transposição poderiam resultar na identificação de novos EGM em genomas de protozoários.

Poucos EGM têm sido identificados e estudados até o momento em tripanossomatídeos, e existem poucos estudos aprofundados sobre sua filogenia e evolução, como o que se refere aos genomas dos Tri-tryps - *Trypanosoma cruzi*, *T. brucei* e *Leishmania major* - relacionados à identificação e caracterização dos seguintes EGM: Ingi (Kimmel *et al*. 1987), Rime (Murphy *et al*. 1987), Sire (Vazquez *et al*. 1994), Viper (Vazquez *et al*. 2000), Tubis (Glauser *et al*. 1994), L1Tc (Martin *et al*. 1995), NARTc (Bringaud *et al*. 2002a; Bringaud *et al*. 2002b) e DIRE (Ghedin *et al*. 2004).

1.3.3.1 Aspectos evolutivos dos EGM

A ampla distribuição dos elementos de transposição sugere que eles tiveram um papel importante na evolução. Uma hipótese é que eles seriam elementos da natureza que auxiliariam na construção do genoma. As habilidades de copiar, transpor e rearranjar outras sequências de DNA, tais como os genes de resistência a antibiótico, podem ser um benefício para os organismos que as possuem. Assim, os elementos de transposição devem ter sido espalhados devido à vantagem seletiva que conferem. Outra hipótese é que os elementos de transposição se espalharam simplesmente porque teriam a habilidade de se multiplicar independentemente dos mecanismos de replicação normal. De acordo com esta hipótese, os elementos de transposição podem ser considerados pouco mais que parasitas genômicos (Orgel & Crick 1980).

As entidades mais representativas dos EGM são os retroelementos que codificam a enzima chave transcriptase reversa. Os retrovírus teriam se desenvolvido a partir de retrotransposons mais simples, através da adição do gene env que sintetiza uma proteína de membrana. Com esta adição, os retrotransposons produziriam uma partícula capaz de escapar de uma célula e entrar em outra. Tal partícula seria infecciosa e forneceria aos retrotransposons a oportunidade de se transporem dentro dos genomas e também entre eles. O contrário também

pode ter ocorrido. Um retrovírus poderia perder sua habilidade de escapar de uma célula e ficar capturado nela. Tal vírus mutante seria reduzido à condição de retrotransposons, capaz de se mover dentro das células, mas não entre elas. O virologista ganhador do Prêmio Nobel Howard Temin (Temin 1995; Temin 1985) descreveu estes cenários contrastantes como um sobe-e-desce nas escalas evolutivas, recorrendo a uma metáfora musical. Os retrotransposons poderiam subir ao nível dos retrovírus e os retrovírus poderiam descer ao nível dos retrotransposons.

1.3.3.2 Retrotransposons

Os retrotransposons ou retroelementos estão muito espalhados e aceita-se que possuam uma origem muito antiga. Devido ao fato de terem uma íntima relação com a transcriptase reversa, é muito provável que tenham surgido a partir do "mundo do RNA". Na origem da vida, todas as biossínteses estiveram presumivelmente mediadas pelo RNA, outra razão para considerar a transcriptase reversa como uma enzima muito antiga (Orgel & Crick 1980).

No "mundo do DNA", o gene ancestral da transcriptase reversa teria tido a oportunidade de replicar-se muitas vezes, contribuindo com a diversidade dos retroelementos. Pelo fato de os retroelementos encontrarem-se mais espalhados que os transposons DNA e as integrases dos retrotransposons/retrovírus serem similares às transposases codificadas por estes transposons DNA (Doak *et al*. 1994), especula-se que os transposons DNA teriam se desenvolvido a partir de antecessores dos retroelementos. Se isso for verdade, o RNA teria sido o antecessor de muitos tipos de DNA. Ultimamente tem sido observado em alguns parasitas que os retroelementos que evoluíram positivamente foram os mais benéficos ou os que conferiram vantagens aos seus hospedeiros (Coffin *et al*. 1997; Doolittle & Sapienza 1980).

A evolução dos eucariotos tem sido marcada por vários processos, os quais incluem: (i) quebra, rejunção e deslocamento de diferentes cromossomos, (ii) duplicações de genes e segmentos, (iii) *shuffling* dos domínios funcionais nos éxons e (iv) conversões de genes. Os retrotransposons não-LTR têm desenvolvido uma história de cerca de 500 a 600 milhões de anos. Eles contêm uma transcriptase reversa similar à encontrada nos íntrons do grupo dos transposons - estes últimos situados nos genomas dos fungos, das plantas e de algumas bactérias (Kempken *et al*. 1998).

1.3.3.3 A evolução dos eucariotos e seus retroelementos

A sequência de eventos necessária na evolução dos eucariotos tem sido debatida exaustivamente e tem sido comprovado que os eucariotos contêm uma mistura de genes descendentes de células ancestrais de Archaeas e Eubactérias (Margulis 1996; Woese *et al*. 1990). O endossimbionte derivado das eubactérias lentamente teria transferido seus genes ao núcleo da célula eucariótica primitiva, chegando assim a depender ainda mais do seu hospedeiro. Este processo de transferência de genes da mitocôndria para o núcleo é funcional em células de leveduras atuais e pode ser observado experimentalmente (Jiang 2002; Ketting *et al*. 1999; Thorsness *et al*. 2002). Através deste processo, os genes da transcriptase reversa presentes nos retroíntrons deveriam ter sido transferidos para o núcleo de maneira passiva e estocástica. A movimentação dos retroíntrons para lugares próximos poderia ter conduzido a uma proliferação de íntrons e à evolução de mecanismos de *splicing* para a remoção dos mesmos (Arkhipova *et al*. 2003; Chapman *et al*. 1992; Gilbert *et al*. 1993; Kempken *et al*. 1998; Logsdon 1998; Nakamura *et al*. 1998a; Simpson *et al*. 2002; Stoltzfus 1994).

Uma diferença interessante entre bactérias e eucariotos que pode estar relacionada com a estabilidade diferencial do RNA é a habilidade dos eucariotos de produzirem proteínas significativamente compridas, como as longas cadeias de poliproteínas codificadas pelos retrotransposons. É importante ressaltar que os genomas de bactérias, na maioria das vezes, codificam proteínas mais curtas do que aquelas codificadas pelos eucariotos, o que está relacionada à economia por compactação do genoma (Margulis 1996; Woese *et al*. 1990).

1.3.3.4 Transcriptase reversa

A origem ancestral da transcriptase reversa tem sido sustentada por dois argumentos principais. O primeiro é teórico e extremamente aceito. Baseia-se na proposta de que o "mundo do RNA" precedeu à formação do "mundo do DNA". Darnell foi o primeiro a defender que a transcriptase reversa deve ter surgido durante o processo da transição entre estes dois mundos e que, portanto, teria que ser considerada como ancestral (Darnell *et al*. 1986). O segundo argumento baseia-se no fato de que os genes da transcriptase reversa estão amplamente distribuídos nos ramos da árvore da vida (Doolittle *et al*. 1989; Eickbush 1994; Malik *et al*. 1999; Xiong & Eickbush 1990).

Além disso, o gene da transcriptase reversa aparentemente apresenta diferentes tipos de rearranjos (Boeke 1997). Além dos retrovírus, existem: (i) os pararretrovírus (que empacotam DNA mas se replicam por transcrição reversa), (ii) as duas maiores classes de retrotransposons (não-LTR e LTR) e (iii) um grupo raro de elementos apresentados em bactérias e genomas de organelas pertencentes aos procariotos. A descoberta das transcriptases reversas em bactérias na forma de DNAms (*short for multicopy single-strand DNA*), retroelementos (Yamanaka K 2002) e retroíntrons (Belfort *et al*. 2002) provém de uma dramática evidência em favor da origem ancestral da transcriptase reversa.

O grupo procariótico inclui três grupos principais de retroelementos:

1) Os rétrons, genes da transcriptase reversa que produzem uma ramificação estrutural chamada DNAms feita por transcrição reversa a partir de um precursor de RNA. Estes rétrons não têm função conhecida nem a habilidade de se movimentarem de forma autônoma (Levin 1995; Yamanaka K 2002),

2) Os retroplasmídeos, encontrados na mitocôndria de alguns fungos e que são replicados por transcrição reversa (Kuiper *et al*. 1988; Walther *et al*. 1999) e

3) Os retroplasmídeos, encontrados na mitocôndria de alguns fungos e que são replicados por transcrição reversa (Zimmerly *et al*. 1995a; Zimmerly *et al*. 1995b).

Argumentos diversos afirmam que as transcriptases reversas do grupo procariótico são ancestrais das transcriptases reversas dos retrotransposons e retrovírus. Por outro lado, existem contra-argumentos para cada uma destas propostas que também são convincentes (Doolittle *et al*. 1989; Eickbush 1994; Eickbush 1997; Nakamura & Cech 1998a; Telesnitsky *et al*. 1997; Wang *et al*. 1993).

1.3.3.5 A distribuição atual da transcriptase reversa

Considerando como premissa que a transcriptase reversa seja uma enzima ancestral, é difícil explicar a distribuição atual de genes da transcriptase reversa nos domínios Eubactéria, Archaea e Eucária. A maioria das espécies de eubactérias sequenciadas (67%) tem perdido genes detectáveis da transcriptase reversa nos genomas (Dai *et al*. 2002a; Dai *et al*. 2002b).

A grande maioria de archaeas perdeu completamente a transcriptase reversa, com exceção da espécie *Methanosarcina sp.*, que possui um genoma muito extenso. Considera-se que na formação desta espécie tenham sido incorporados grandes segmentos de genomas de eubactérias

por meio de um evento de transferência lateral tardia (Deppenmeier *et al*. 2002). Estas archaeas contêm um grupo de retroíntrons similares aos encontrados em eubactérias (Dai & Zimmerly 2002a; Dai & Zimmerly 2002b).

Em contraste, os genes da transcriptase reversa são encontrados virtualmente em todos os genomas eucarióticos, contendo de 20 a mais de 500 mil cópias por genoma. Estas transcriptases reversas pertenceriam aos retrotransposons não-LTR ou LTR.

Bushman, em 2002, poeticamente descreveu os genomas eucarióticos como "genes flutuando em um mar de retrotransposons". Dentre os retrotransposons mais conhecidos que exemplificam esta frase estão os retrotransposons não-LTR, os SINE e os retrovírus endógenos do genoma humano, os quais chegam a, aproximadamente, um milhão de cópias (Smit *et al*. 1996). Outro exemplo é o genoma do milho. Estima-se que contenha cerca de 200 mil cópias de retrotransposons intactos (SanMiguel *et al*. 1996). A dramática discrepância na distribuição dos retroelementos entre procariotos e eucariotos sugere fortemente que exista alguma característica especial que permita que os eucariotos apresentem um estado permissivo para a atividade da transcriptase reversa, devido ao grau de proliferação e adaptação dos retrotransposons nestes genomas.

1.3.3.6 Tripanossomatídeos

A posição sistemática do gênero *Trypanosoma* tem variado nos últimos anos em relação à posição das taxas superiores. De acordo com a última classificação adotada pela Sociedade de Protozoologia (Levine *et al*. 1980), a superclasse Mastigophora passou a Subfilo, o Subfilo Sarcomastigophora passou a Filo e o Protozoa passou a Subreino.

A posição sistemática dos tripanossomatídeos é a seguinte (Levine *et al*. 1980):

Reino: Protista

Subreino: Protozoa

Filo: Sarcomastigophora

Subfilo: Mastigophora

Classe: Zoomastigophorea

Ordem: Kinetoplastida

Subordem: Trypanosomatina

Família: Trypanosomatidae

Gênero 1: *Leptomonas*

Gênero 2: *Herpetomonas*

Gênero 3: *Crithidia*

Gênero 4: *Blastocrithidia*

Gênero 5: *Phytomonas*

Gênero 6: *Leishmania*

Gênero 7: *Endotrypanum*

Gênero 8: *Trypanosoma*.

Os quatro primeiros gêneros da família Trypanosomatidae são parasitas monogenéticos de animais invertebrados (especialmente insetos), ou seja, têm um só tipo de hospedeiro. No entanto, os quatro últimos são parasitas digenéticos, pois mantêm o ciclo de vida alternado entre dois hospedeiros: um invertebrado, que geralmente é um hospedeiro intermediário ou vetor e o outro, um animal vertebrado ou uma planta. Estes parasitas são encontrados na corrente sanguínea ou nos tecidos (*Leishmania*, *Endotrypanum*, *Trypanosoma*) e no látex ou em outro fluído (*Phytomonas*).

O gênero *Trypanosoma* é dividido em duas seções: Stercoraria e Salivaria. A seção Stercoraria inclui espécies (*T. cruzi*, *T. lewisi* e *T. theileri*) que se desenvolvem no intestino posterior de seus vetores, enquanto as espécies pertencentes à seção Salivaria se desenvolvem no intestino anterior e são transmitidas via saliva dos vetores (*T. brucei*, *T. vivax* e *T. congolense*) (Hoare 1966; Vickerman 1976).

Os protozoários patogênicos *L. major*, *T. cruzi* e *T. brucei* (família Trypanosomatidae, ordem Kinetoplastida) causam doenças em milhões de humanos e infecções em outros mamíferos, especialmente em países de regiões tropicais e subtropicais (Barrett *et al*. 2003; Barrett *et al*. 2002). Estes três são referidos como os Tri-tryps e compartilham algumas características: (i) possuem estruturas subcelulares como kinetoplastos e glicossomos, (ii) são transmitidos por diferentes insetos e (iii) possuem diferentes tecidos-alvos e diferentes mecanismos de patogenicidade em seus hospedeiros mamíferos (De Souza 2002). O genoma da espécie *T. cruzi* foi publicado em 2005 (El-Sayed *et al*. 2005a), juntamente com o da *T. brucei* (Berriman *et al*. 2005) e *L. major* (Ivens *et al*. 2005).

Várias outras espécies pertencentes à família Trypanosomatidae estão tendo os seus genomas completa ou parcialmente sequenciados, entre elas *T. vivax*

(http://www.sanger.ac.uk/Projects/T_vivax/) (Davila *et al*. 2003) e *T. congolense* (http://www.sanger.ac.uk/Projects/T_congolense/) (Hertz-Fowler *et al*. 2004). Causadoras de perdas econômicas à indústria de animais domésticos na África e na América do Sul, estas duas espécies são os hemoflagelados mais importantes de interesse veterinário. Enquanto a espécie *T. congolense* é restrita à África, a *T. vivax* foi diagnosticada nos dois continentes. Ambas são transmitidas pela mosca tsé-tsé (*Glossina sp.*) na África, enquanto que, na América do Sul, a *T. vivax* (aparentemente) só é transmitida mecanicamente por moscas hematófagas (Hertz-Fowler & Berriman 2004). *L. braziliensis*, *L. infantum*, *L. mexicana*, *L. seymouri,* assim como outros protozoários dos gêneros *Babesia*, *Bodo*, *Crithidia*, *Dictyostelium*, *Eimeria*, *Entamoeba*, *Leptomonas*, *Neospora*, *Plasmodium*, *Theileria* e *Toxoplasma* também estão sendo sequenciados e as informações podem ser obtidas na página do Instituto Sanger (http://www.sanger.ac.uk/Projects/).

1.3.3.7 Os EGM em protozoários

A maioria dos eucariotos analisados contém ao menos uma família dos EGM (Kazazian 2004). Em contraste, cinco de quinze genomas de eucariotos unicelulares sequenciados na atualidade (http://genomesonline.org/) perderam os EGM, entre eles o parasita intracelular *Encephalitozoon cuniculi* (Katinka *et al*. 2001) e quatro membros dos protozoários patógenos Apicomplexa. São eles: *Plasmodium falciparum* (Gardner *et al*. 2002), *P. yoelii yoelii* (Carlton *et al*. 2002), *Crystosporidium hominis* (Xu *et al*. 2004) e *C. parvum* (Abrahamsen *et al*. 2004). Isto sugere que uma taxa significativa de eucariotos unicelulares pode ter perdido EGM ativos. Entretanto, desde que nenhum destes cinco genomas contenha vestígios detectáveis de EGM, também não se pode descartar a hipótese de que esses genomas nunca contiveram EGM.

Ghedin *et al*. (2004) comparou as espécies *T. brucei*, *T. cruzi* e *L. major* usando a informação do sequenciamento dos seus genomas e comparou as sequências dos seus grandes fragmentos cromossômicos, os quais, apesar do alto grau de divergência no nível de sequência, exibem uma notável conservação na ordem dos genes. Isto sugere que a seleção tem mantido a ordem dos genes nos tripanossomatídeos por centenas de milhões de anos de evolução. Os poucos sítios de rearranjos nos genomas nessas espécies são marcados pela presença de "elementos similares aos retrotransposons", sugerindo que possam ter exercido um importante papel na organização e evolução dos genomas de tripanossomatídeos. Um retroelemento

degenerado foi identificado na espécie *L. major* e considerado o primeiro dos elementos encontrados, o que sugere que os retroelementos estariam presentes em um antecessor comum a todas as espécies de tripanossomatídeos.

Todos os retrotransposons não-LTR encontrados em tripanossomatídeos analisados até agora (Berriman *et al*. 2005; El-Sayed *et al*. 2005a; El-Sayed *et al*. 2005b) são classificados nos grupos CRE e *ingi*. O grupo CRE está composto pelos seguintes retrotransposons sítio-específicos: SLACS em *T. brucei* (Aksoy *et al*. 1990), CZAR em *T. cruzi* (Villanueva *et al*. 1991) e CRE1/CRE2 em *Crithidia fasciculata* (Gabriel *et al*. 1990). Os retrotransposons *ingi* em *T. brucei* e L1*Tc* em *T. cruzi* - do grupo *ingi* - estão dispersos no genoma do hospedeiro (Kimmel *et al*. 1987; Martin *et al*. 1995; Murphy *et al*. 1987), embora apresentem um sítio específico relativo para a inserção (Bringaud *et al*. 2004). Estas análises mostram que *L. major* tem eliminado todos os retrotransposons não-LTR ativos presentes no seu tripanossomatídeo ancestral, enquanto os genomas do gênero *Trypanosoma* ainda contêm elementos ativos.

Por outro lado, as regiões de sintenia presentes em *T. cruzi* foram interrompidas por um conjunto de genes que pertencem à família multigene RHS (Ghedin *et al*. 2004). Os genes RHS - descritos originalmente em *T. brucei* - possuem *hot spots* para as inserções dos retroelementos (Bringaud *et al*. 2002a; Bringaud *et al*. 2002b) e geralmente estão situados próximos aos telômeros cromossômicos (El-Sayed *et al*. 2003; Hall *et al*. 2003). Não existe evidência da existência destes genes RHS nas terminações teloméricas dos cromossomos em *T. cruzi*. Diferentes regiões subteloméricas em *T. cruzi* têm apresentado cópias de um elemento similar ao retroposon SIRE (Chiurillo *et al*. 1999). Outro retrotransposon LTR originalmente caracterizado no genoma da espécie *T. cruzi* foi o chamado VIPER.

Como relatado em plantas e mamíferos, os não-LTR estão distribuídos randomicamente e constituem os EGM mais abundantes descritos no genoma de *T. brucei* (*ingi* e RIME) e *T. cruzi* (L1*Tc*) (Hasan *et al*. 1984; Kimmel *et al*. 1987; Martin *et al*. 1995; Murphy *et al*. 1987), e o número de cópias por genoma dos retrotransposons L1*Tc*/ NAR*Tc* em *T. cruzi* está na mesma escala que os o número de cópias por genoma dos retrotransposons *Ingi*/RIME em *T. brucei*/TRU-927 (Bringaud *et al*. 2002a; Bringaud *et al*. 2002b; Okada *et al*. 1997).

1.4 A telomerase e os elementos de retrotransposição em Tri-tryps baseados em análises filogenéticas dos domínios da transcriptase reversa

1.4.1 A relevância dos transposons para os estudos de evolução do genoma

Os EGM são sequências repetitivas atualmente reconhecidas pelas suas diversas funções evolutivas e estão divididos nas classes I ou retroelementos e II ou transposons DNA (Finnegan 1989). A classe I utiliza um processo conhecido como retrotransposição e inclui os retrotransposons LTR e não-LTR (Craig *et al*. 2002; Kajikawa *et al*. 2002; Malik *et al*. 2000; Prak *et al*. 2000). Estes elementos transponíveis são de interesse para geneticistas (como uma ferramenta experimental), para anotadores de genomas (que os consideram geralmente como DNA lixo para separá-los durante o processo de anotação) e para os biólogos evolutivos e estruturais (por muitas razões) (Holmes 2002).

As seguintes perguntas estão relacionadas ao interesse evolutivo e significativo dos transposons: Como os elementos transponíveis evoluem no nível de sequência? Qual a informação contida nas taxas de transposição e recolonização? Qual a informação evolutiva e filogenética da presença dos transposons em espécies estreitamente relacionadas? Como a transcriptase reversa e a transposase poderiam ser usadas para reconstruir a vida evolutiva dos EGM em Tri-tryps e protozoários? Quais enzimas ou domínios de enzimas são necessários para obter uma reconstrução válida? É possível o desenho de marcadores moleculares úteis a partir dos EGM para os tripanossomatídeos patogênicos?

O papel das sequências que sofrem eventos de repetição, transposição e duplicação na evolução, é considerado um tema interessante e polêmico (Doolittle & Sapienza 1980; McClintock 1984; Orgel & Crick 1980). As sequências repetitivas são muito numerosas, o que torna sua anotação um problema. Este é particularmente o caso das famílias de sequências repetitivas, que muitas vezes carregam seus próprios genes (por exemplo, a transposase e a transcriptase reversa), o que pode confundir sua anotação em larga escala. Muitos programas são utilizados atualmente para a identificação e anotação dos EGM. O RECON, por exemplo, é uma

ferramenta utilizada para a anotação automática de transposons e DNA repetitivos (Bao *et al*. 2002).

1.4.2 Evolução dos telômeros: uma perspectiva

A telomerase - uma ribonucleoproteína muitas vezes considerada filogeneticamente como uma transcriptase reversa - é a enzima responsável pela replicação das extremidades dos cromossomos - telômeros - na maioria dos eucariotos (Greider *et al*. 1987). A subunidade catalítica da telomerase foi identificada em *Euplotes aediculatus* e *Saccharomyces cerevisiae* (Lingner *et al*. 1997a; Lingner *et al*. 1997b) e, posteriormente em *Schizosaccharomyces* (Nakamura *et al*. 1997) e em humanos (Harrington *et al*. 1997; Kilian *et al*. 1997; Meyerson *et al*. 1997; Nakamura & Cech 1998a; Nakamura *et al*. 1997; Nakayama *et al*. 1997). A primeira constatação de que os telômeros possuem uma importância vital nas funções biológicas foram expostos nos trabalhos pioneiros de McClintock e Muller em *Drosophila* (Herskowitz *et al*. 1954) e milho (McClintock 1941a; McClintock 1938; McClintock 1941b).

A atenção do público em geral tem sido estimulada através das conexões definidas entre os telômeros, o envelhecimento e o câncer. Novas técnicas foram desenvolvidas com o intuito de melhor compreender a estrutura, a função e a rápida evolução dos telômeros. Os avanços foram obtidos a partir do estudo de vários organismos modelos (leveduras, protozoários, insetos, vertebrados e plantas), permitindo desta maneira a comparação entre eles e outras espécies diferentes, inferindo assim filogenia e evolução de uma perspectiva mais consistente e precisa (Fajkus *et al*. 2005).

É provável que a telomerase seja muito antiga, uma vez que é encontrada em eucariotos ancestrais que representam a origem de muitas das principais linhagens (ciliados, animais, fungos, plantas verdes). A perda da telomerase seria um evento catastrófico se não houvesse uma imediata substituição por sistemas alternativos como a retrotransposição, a recombinação e/ou a replicação (Louis 2002).

1.4.3 Qual é a relação entre os motivos da transcriptase reversa e a estrutura da subunidade catalítica da telomerase? Existe um contexto evolutivo?

Encontrar domínios da transcriptase reversa na subunidade catalítica da telomerase não foi algo inesperado. A telomerase polimeriza o DNA usando um modelo de RNA e é, portanto e por definição, uma transcriptase reversa, embora a relação evolutiva com os diferentes tipos de transcriptases reversas seja ainda desconhecida (Nakamura & Cech 1998a). Hoje, todos os sete domínios previamente definidos da transcriptase reversa (Eickbush 1999; Eickbush 1992) foram identificados na telomerase (Nakamura *et al*. 1997), e a importância de alguns deles em sua atividade demonstrou-se por mutagênese em *S. cerevisiae* (Lingner *et al*. 1997a; Lingner *et al*. 1997b) e em humanos (Harrington *et al*. 1997; Nakamura *et al*. 1998b; Weinrich *et al*. 1997).

O uso da telomerase é tão difundido em eucariotos que a mais simples explicação seria que a manutenção dos telômeros é antiga, coincidindo ou mesmo precedendo às primeiras células eucarióticas. Há muitos problemas para determinar a antiguidade da telomerase. As telomerases infelizmente variam muito de tamanho e sua similaridade de sequências é demasiado baixa para ser útil na análise filogenética, exceto entre espécies estreitamente relacionadas. Ainda não há um antepassado de RNA para a telomerase, mas já foram identificados domínios essenciais da transcriptase reversa para este gene, o que tornou possível localizar a telomerase no contexto de outras transcriptases reversas pela análise da sequência, obtendo assim algumas pistas sobre sua origem (Nakamura & Cech 1998a).

1.4.4 A telomerase nos Tri-tryps

A maioria dos estudos da enzimologia da telomerase tem sido feita em ciliados, mamíferos ou sistemas modelo de levedura. Todos estes organismos representam ramos recentes da árvore filogenética de eucariotos se comparados às espécies de parasitas protozoários Kinetoplastida (Sogin 1991). A telomerase foi identificada em numerosos organismos, mas não estudada extensamente em protozoários. Há várias questões relevantes para a consideração da manutenção dos telômeros dependentes da telomerase em tripanossomatídeos. Estes parasitas possuem ciclos de vida com diferentes formas de replicação e infecção específicas para um determinado inseto ou animal hospedeiro. A manutenção do comprimento dos telômeros no

parasita poderia, portanto, estar limitada a um determinado estágio de desenvolvimento. Além disso, apesar da divergência evolutiva entre parasita e hospedeiro, as enzimas telomerases de ambos sintetizam a mesma sequência de repetição telomérica (Munoz *et al*. 2004).

1.5 ProtozoaDB: A visualização dinâmica e exploração dos genomas dos protozoários

O desenvolvimento de ferramentas para a análise filogenética e os bancos de dados têm um enorme impacto no tratamento de dados complexos e no volume das informações processadas. O número de estudos filogenéticos está aumentando rapidamente, e o grande volume de dados tornou-se um problema de pesquisa importante quando o assunto é sua análise.

O processo filogenético está representado pelo uso de diferentes programas e algoritmos executados ao mesmo tempo ou de maneira consecutiva automática. Os processos vão desde a construção do alinhamento múltiplo de sequências, passando pela geração da matriz evolutiva e pela definição dos melhores parâmetros/algoritmos dos programas filogenético-evolutivos, até as etapas da execução na criação da árvore, que variam de estudo para estudo, dependendo das características biológicas de cada sequência de dados - genes, enzimas.

O acesso sistemático das ferramentas desse processo corresponde às etapas de um fluxo de trabalho filogenético, onde diferentes parâmetros/algoritmos geram diferentes árvores. A escolha da melhor árvore é um passo difícil e leva tempo, devido à comparação das diferentes topologias e valores de consistência. É importante que o pesquisador biólogo possa usar, no mesmo estudo, diferentes parâmetros e programas e que tenha a opção de armazenar estes resultados para uma análise futura mais detalhada. Este recurso propõe a proveniência de dados e o incremento da qualidade do estudo filogenético.

O armazenamento de uma árvore filogenética no formato NEXUS geralmente é feito em um único campo ou variável. Esta abordagem não permite realizar análises complexas sobre a árvore devido ao fato de este formato armazenado não ser flexível ou suficiente. Existe uma técnica que se baseia na decomposição do padrão de uma árvore - formato NEWICK, extraído do NEXUS - em um conjunto de caminhos - pais e filhos - para tentar marcar o escore de reconstrução da árvore no banco de dados. A partir daí, podem ser feitas consultas baseadas em perguntas de interesse biológico que envolvem, por exemplo, conhecer o ancestral comum de um nó pertencente a uma árvore em estudo, obter as agrupações de todos os nós ou filhos dado um

valor de *bootstrap* ou procurar por uma espécie nas diferentes árvores, cujo clado seja suportado por um valor de *bootstrap* maior que 80. No entanto, este método não pode ser utilizado para realizar consultas cuja entrada não contenha informações estruturais.

Armazenar a árvore filogenética de uma forma flexível no banco de dados corresponde a relacionar os ramos (táxons, espécies e gêneros, entre outros) da árvore entre si ou entre valores de suporte. Assim, é possível aumentar a complexidade das questões estruturais que estarão disponíveis para que o biólogo gere relatórios sobre os organismos em diferentes árvores filogenéticas.

O objetivo do presente trabalho foi fornecer um sistema integrado composto por três componentes: (i) uma interface *web* para inserir dados, parâmetros e criar ou armazenar consultas, (ii) uma ferramenta de fluxo de trabalho flexível para orquestrar as etapas do processo filogenético e (iii) um banco de dados projetado especificamente para permitir consultas complexas e armazenar os dados de fluxo de trabalho de proveniência. Este trabalho propõe uma solução para o tratamento e armazenamento através do desenvolvimento de um módulo filogenético para o banco de dados genômicos denominado GUS (*Genomic Unified Schema*) e através de serviços *web* para o tratamento dos dados filogenéticos.

CAPÍTULO 2 - OBJETIVOS

1- Geral

Inferir a filogenia e a evolução dos protozoários usando uma abordagem filogenômica.

2- Específicos

(i) Analisar e inferir a filogenia dos genes relacionados à resistência a drogas em genomas de protozoários,

(ii) Reconstruir a árvore de espécies de protozoários,

(iii) Realizar estudos de filogenômica nos Elementos Genéticos Móveis (EGM) em protozoários,

(iv) Inferir a filogenia da telomerase e dos elementos de retrotransposição em Tri-tryps,

(v) Adaptar e ampliar o esquema Phylo ao banco de dados GUS para o armazenamento da informação filogenética.

CAPÍTULO 3 - MATERIAIS E MÉTODOS

3.1 Reconstrução automática de análises filogenômicas (ARPA) de genes relacionados à resistência a drogas em genomas de protozoários

3.1.1 Projeto ARPA

3.1.1.1 Execuções do ARPA

Este sistema é composto por 18 programas de bioinformática e filogenia estruturados de forma que permite uma execução automática, isto é sem necessidade de interferência do usuário. O ARPA pode ser executado via interface *web* ou localmente via linha de comando através de um terminal. A tabela 3.1 apresenta todos os programas utilizados no sistema bem com sua respectiva função.

O ARPA alinha, limpa e edita as sequências dos arquivos multifasta e as converte para os formatos phylip e nexus, utilizando o programa READSEQ. Em seguida, faz a procura do modelo evolutivo baseado nos critérios AIC (*Akaike information criterion*), usando o programa MODELGENERATOR (Keane *et al*. 2006).

Obs: O alinhamento é feito pelo ARPA com o programa MAFFT v6.240 (Katoh *et al*. 2005; Katoh *et al*. 2002; Katoh *et al*. 2008) e as árvores filogenéticas são feitas com os programas disponíveis na tabela 3.4 (figura 3.1).

3.1.1.2 Disponibilidade

A interface gráfica *web* do ARPA foi desenvolvida como parte de uma solução de *software* livre. Sua estrutura usa o sistema operacional Linux Ubuntu 8.04 e o servidor *web* Apache 1.3. A interface *web* utiliza HTML e JavaScript ('Active script' in IE), os quais são executados em Internet Explorer (IE) 6 e Firefox 1.5 ou em versões mais recentes. O JavaScript 1.2 foi usado como linguagem para as validações na máquina do cliente e a tecnologia Ajax (lib effects.js) para uma interface amigável.

O sistema ARPA assim como sua interface *web* estão hospedados na plataforma de Bioinformática do Instituto Oswaldo Cruz - FIOCRUZ e podem ser acessados no endereço

http://arpa.biowebdb.org. Eles estão sendo integrados aos sistemas de banco de dados ProtozoaDB (http://protozoadb.biowebdb.org) e Stingray (http://stingray.biowebdb.org/).

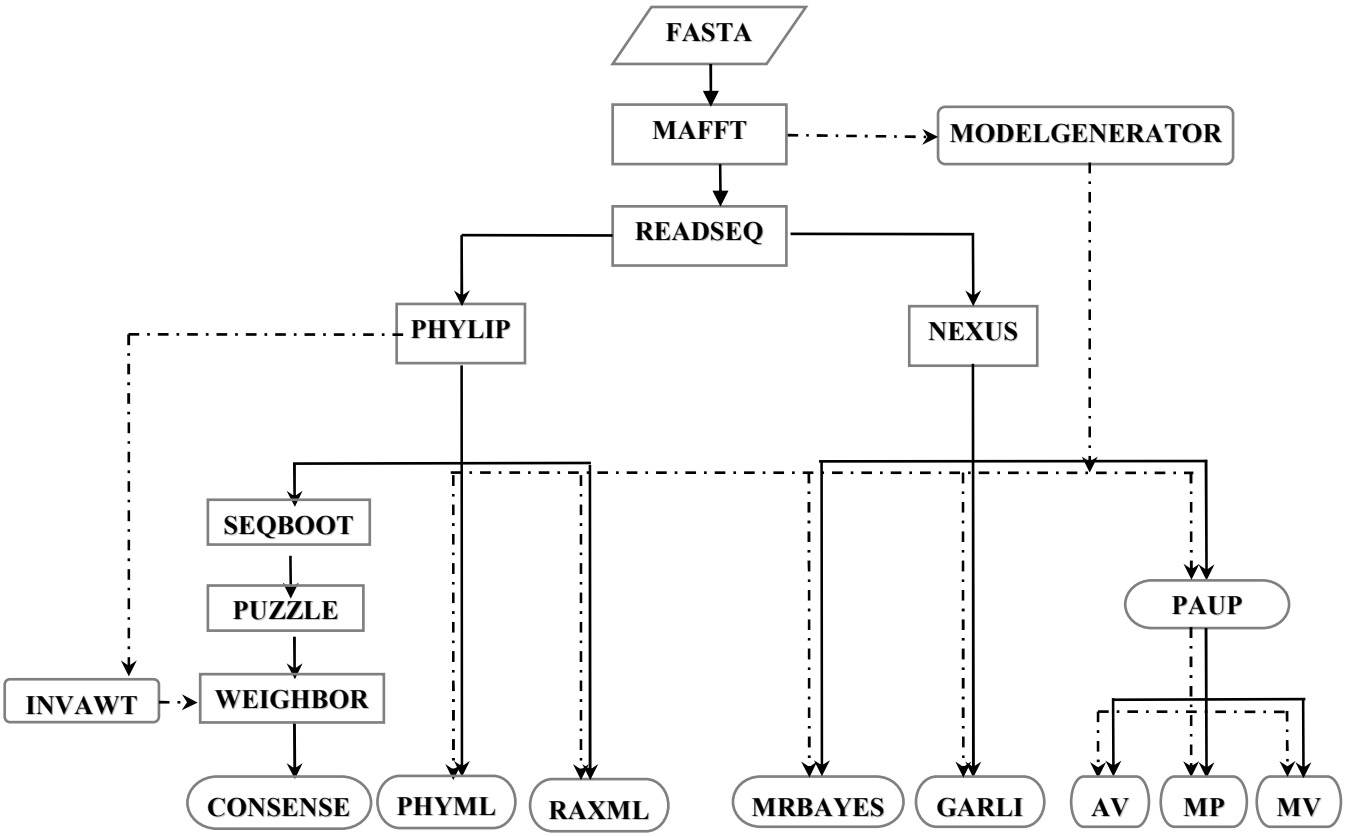


Figura 3.1 - Fluxograma do sistema ARPA

O ARPA está formado por 18 programas bioinformática e foi automatizado para alinhar, eleger o melhor modelo evolutivo, limpar e formatar os arquivos de entrada e construir as árvores filogenéticas.

Tabela 3.1 - Programas implementados no sistema ARPA

Programas	Função	Versão	Referência	Endereço *web*
ALIGN-M	Alinhamento múltiplo	6.240	(Van Walle *et al*. 2004)	http://bioinformatics.vub.ac.be
CLUSTALW	Alinhamento múltiplo	1.83	(Larkin *et al*. 2007; Thompson *et al*. 2002; Thompson *et al*. 1994)	http://www.ebi.ac.uk/Tools/clustalw2/index.html
MUSCLE	Alinhamento múltiplo	3.6	(Edgar 2004)	http://www.drive5.com/muscle
MAFFT	Alinhamento múltiplo	6.240	(Katoh *et al*. 2005; Katoh *et al*. 2002; Katoh & Toh 2008)	http://align.bmr.kyushu-u.ac.jp/mafft/software/
PROBCONS	Alinhamento múltiplo	1.12	(Do *et al*. 2005)	http://probcons.stanford.edu/
T-COFFEE	Alinhamento múltiplo	8.14	(Notredame *et al*. 2000)	http://www.ebi.ac.uk/Tools/t-coffee/index.html
MODELGENERATOR	Eleição do modelo evolutivo	0.85	(Keane *et al*. 2006)	http://bioinf.may.ie/software/modelgenerator/
READSEQ	Conversão de formato dos alinhamentos	2.1.26	(Gilbert 2003)	http://itc.virginia.edu/achs/molbio/phylip/readseq.html
GBLOCKS	Extração de regiões conservadas	0.91b	(Castresana 2000; Talavera *et al*. 2007)	http://molevol.cmima.csic.es/castresana/Gblocks.html
TRIMAL	Extração de regiões conservadas	1.2rev59	(Capella-Gutierrez *et al*. 2009)	http://trimal.cgenomics.org
GARLI	Análise filogenética	0.96 Beta	(Zwickl 2006)	http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html
MRBAYES	Análise filogenética	3.1.2	(Huelsenbeck *et al*. 2001; Ronquist *et al*. 2003)	http://mrbayes.csit.fsu.edu/
PHYLIP	Análise filogenética	3.66	(Felsenstein 2005)	http://evolution.gs.washington.edu/phylip.html
PAUP	Análise filogenética	4.0b10	(Swofford 2002; Wilgenbusch *et al*. 2003)	http://paup.csit.fsu.edu/index.html
PHYML	Análise filogenética	2.4.4	(Guindon *et al*. 2003; Guindon *et al*. 2005)	http://atgc.lirmm.fr/phyml/binaries.html
TREE-PUZZLE	Análise filogenética	5.2	(Schmidt *et al*. 2002; Schmidt *et al*. 2007),	http://www.tree-puzzle.de/
PUZZLEBOOT	Análise filogenética	5.2	(Schmidt *et al*. 2002; Schmidt & von Haeseler 2007)	http://www.tree-puzzle.de/#puzzleboot
RAXML	Análise filogenética	7.2.1	(Stamatakis *et al*. 2005)	http://wwwkramer.in.tum.de/exelixis/software.html
WEIGHBOR	Análise filogenética	1.2.1	(Bruno *et al*. 2000)	http://www.t6.lanl.gov/billb/weighbor/

3.1.2 Seleção e preparação dos genes de resistência a drogas em protozoários e análises filogenéticas

O grupo de sequências de genes de resistência a drogas (tabela 3.2) preditas e anotadas que foram usadas neste estudo foi obtido dos seguintes bancos de dados (figura 3.2, tabela 3.3, anexo 1, 2, 7, 8):

- REFSEQ *Release35.catalog* - 01/11/2009 (http://www.ncbi.nlm.nih.gov/RefSeq/),

- GENBANK *NCBI-Flat File Release 172.0* (http://www.ncbi.nlm.nih.gov/Genbank/),

- COG 17/04/2003 e KOG 21/07/2003 (http://www.ncbi.nlm.nih.gov/COG/new/) e

Para cada grupo de sequência de genes de resistência a drogas, foram criados e calibrados perfis *hidden Markov models* (HMM) (Eddy 1998) usando o pacote HMMER versão 2.3.2 com os programas hmmbuild e hmmcalibrate.

- PFAM Pfam23.0 (http://pfam.sanger.ac.uk/) - deste banco de dados foram obtidos os perfis HMM - já construídos e calibrados com o HMMER2.0 [2.3.2] - para cada um dos cinco grupos de genes.

Cada um destes perfis HMM foi comparado aos genomas de protozoários catalogados e disponíveis nos bancos de dados REFSEQ e GENBANK, utilizando o HMMER – programa hmmpfam com *E-value* de 10^{-5} (tabela 3.3).

Tabela 3.2 - Genes de resistência a drogas usados para a filogenia

Gene	Abreviatura
Transportador ABC-MRPA (*Drug Transport Multidrug Resistance Associated Protein A*)	MRPA
Aquaporina	AQP
Tripanotiona Redutase	TRYR
Glicoproteína de Superfície de 63kDa	GP63
Proteínas de Choque Térmico de 70kDa	hsp70

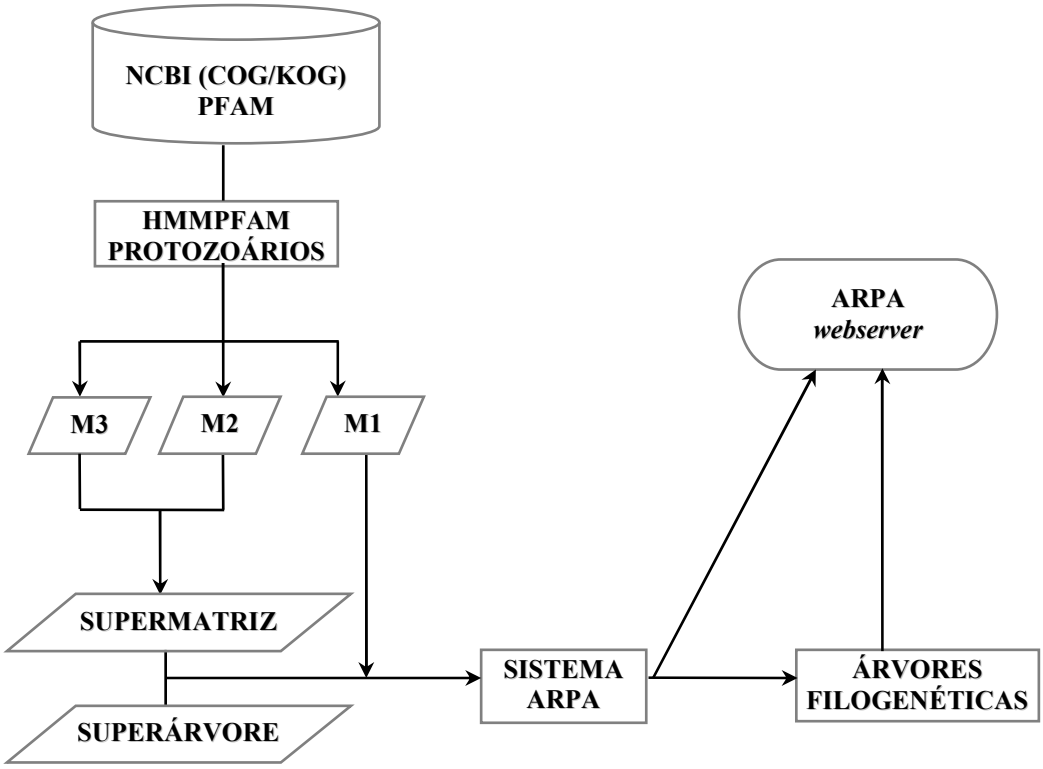


Figura 3.2 - Fluxograma utilizado para construir o banco de genes de candidatos de resistência às drogas

As árvores filogenéticas individuais, a árvore da supermatriz e a superárvore foram construídas com o ARPA.

Tabela 3.3 - Bancos de dados utilizados para a obtenção das sequências dos genes de resistência a drogas e as sequências dos genomas de protozoários usadas nas análises de comparação por similaridade

Banco de dados	Descrição	Tipo	Número de entradas	Versão	Endereço *web*
COG	Sistema de classificação baseado nas relações entre os genes ortólogos, Este sistema representa uma análise de genômica comparativa que facilita tanto a anotação funcional de genomas em grande escala, como os estudos evolutivos.	Proteína	4.873	17/04/2003	http://www.ncbi.nlm.nih.gov/COG/
KOG	Versão do COG, para sete genomas eucarióticos completos: *S. cerevisiae*, *S, pombe*, *E. cuniculi*, *A. thaliana*, *C. elegans*, *D. melanogaster* e *H. sapiens*.	Proteína	4.852	21/07/2003	http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi
GENBANK	Banco de dados que incorpora sequências de DNA de mais de 105.000 organismos diferentes, principalmente através da submissão direta de dados e as disponibiliza ao público.	Proteína / nucleotídeo	13,5 bilhões de bases de nucleotídeos de 12,8 milhões de sequências diferentes	*NCBI-Flat File Release 172.0* - 06/15/2009	http://www.ncbi.nlm.nih.gov/Genbank/
PFAM	O banco de dados PFAM é uma grande coleção de famílias de proteínas, cada uma representada por alinhamentos de sequência múltiplos e perfis HMM	Proteína	5.323.441 sequências, 1.738.474.641 resíduos e 10.340 famílias	23.0	(http://pfam.sanger.ac.uk/)
REFSEQ	Banco de dados de proteínas não redundantes do NCBI	Proteína	9.662.677	*Release35. catalog* - 01/11/2009	http://www.ncbi.nlm.nih.gov/RefSeq/

3.1.2.1 Metodologia 1 (M1)

A metodologia M1 apresenta-se esquematizada na tabela 3.4.

Todos os *hits* do hmmpfam dos Genes de Resistência a Drogas em Protozoários foram usados na reconstrução filogenética executando o sistema ARPA (http://arpa.biowebdb.org) via linha de comando. Todos os *hits* do hmmpfam referem-se a todos os *hits* resultantes da comparação com o HMMER - programa hmmpfam que pertencem a todos os táxons dos protozoários encontrados nos cinco genes em estudo.

O programa MAFFT v6.240 foi usado para a construção dos alinhamentos e o programa PAUP 4.0b10, utilizando o algoritmo AV com *bootstrap* de 10.000, foi usado para a construção das árvores.

3.1.2.2 Metodologia 2 (M2)

A metodologia M2 apresenta-se esquematizada na tabela 3.4.

Todos os melhores *hits* do hmmpfam dos Genes de Resistência a Drogas em Protozoários foram usados na reconstrução filogenética executando o sistema ARPA. Os melhores *hits* do hmmpfam referem-se aos primeiros *hits* resultantes da comparação com o HMMER - programa hmmpfam que pertencem, além disso, a um exemplar por espécie de cada um dos táxons dos protozoários encontrados nos cinco genes em estudo.

(M2-1): Foram formados dois grupos considerando:

(i) espécies apresentando o melhor *hit* em todos os genes e

(ii) espécies apresentando o melhor *hit* em quatro genes.

(M2-2): Para os dois grupos gerados a partir da M2-1 (i e ii), foram usados três diferentes tipos de execuções:

(A) sequências de genes completos,

(B) sequências de genes trimados com o programa GBLOCKS 0.91b (Castresana 2000; Talavera & Castresana 2007) e

(C) sequências de genes trimados com o programa TRIMAL v1.2 (http://bioinfo.cipf.es/trimal/) (Capella-Gutierrez *et al*. 2009) (figura 3.3).

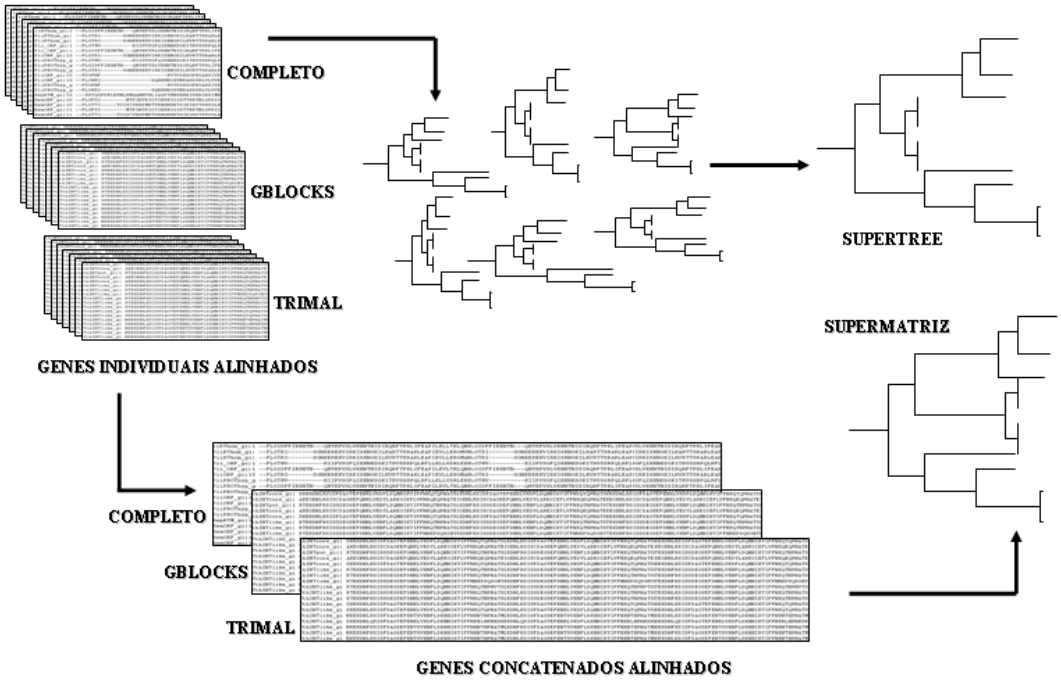


Figura 3.3 - Fluxograma utilizado na construção da árvore da supermatriz e da superárvore

As árvores filogenéticas individuais, a árvore da supermatriz e a superárvore foram construídas com o ARPA.

Desta maneira foram obtidos os grupos de sequências **iA, iB, iC, iiA, iiB, iiC** (figura 3.4):

(iA) espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos,

(iB) espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o GBLOCKS 0.91b,

(iC) espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o TRIMAL v1.2,

(iiA) espécies apresentando o melhor *hit* em quatro genes / sequências de genes completos,

(iiB) espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o GBLOCKS 0.91b,

(iiC) espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o TRIMAL v1.2.

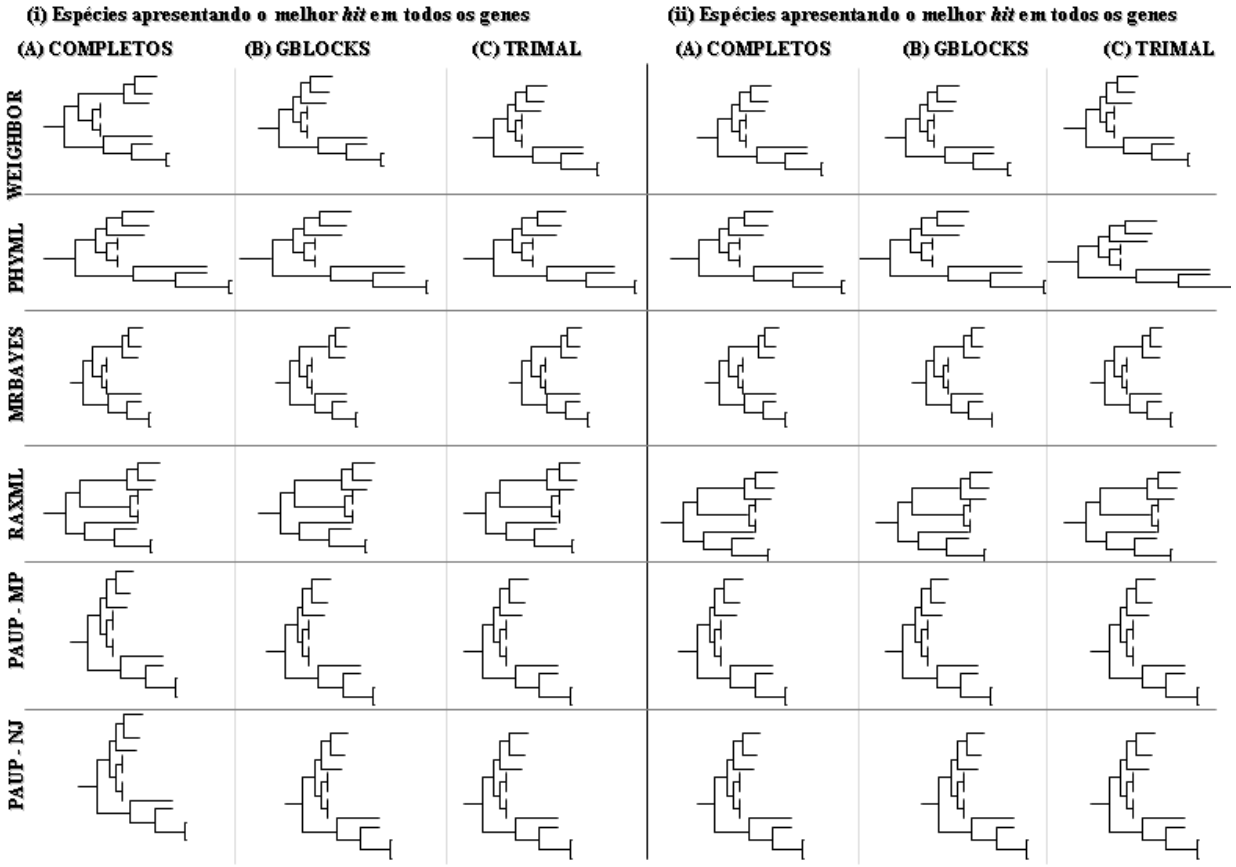


Figura 3.4 - Esquema do total das árvores filogenéticas individuais construídas com os diferentes programas de filogenia

O esquema apresenta as diferentes árvores filogenéticas individuais segundo a Metodologia M2 para o conjunto de genes de candidatos de resistência às drogas.

3.1.2.3 Metodologia 3 (M3)

A metodologia M3 apresenta-se esquematizada na tabela 3.4.

Foram construídas:

- árvores filogenéticas individuais,

- uma única árvore da supermatriz e

- uma única superárvore,

Todas as árvores foram construídas com o ARPA e editadas e visualizadas com o programa MEGA 4 (Kumar *et al*. 2008) (figura 3.4).

(M3-1): As árvores filogenéticas individuais (Grupo iA, iB, iC, iiA, iiB, iiC) foram construídas com o ARPA utilizando os programas: PAUP-AV 4.0b10, PAUP-MP 4.0b10, PHYML 2.4.4, RAXML 7.2.1, MRBAYES 3.1.2 e o fluxograma de execução do WEIGHBOR (que inclui o programa SEQBOOT do pacote PHYLIP 3.66 (Felsenstein 1989), os programas TREE-PUZZLE 5.2 (Schmidt *et al*. 2002; Schmidt & von Haeseler 2007) e WEIGHBOR 1.2.1 (Bruno *et al*. 2000) e o programa CONSENSE do pacote PHYLIP 3.66).

A matriz do modelo evolutivo WAG foi usada com os algoritmos MV e IB, com exceção do gene hsp70, para o qual foi usada a matriz RtREV.

O grupo iC foi considerado o mais informativo e foi usado na construção das árvores abaixo:

(M3-2): A árvore da supermatriz - que usou os genes concatenados alinhados por um programa em perl - foi construída com o PAUP 4.0b10 - algoritmo AV - *bootstrap* de 10.000.

(M3-3): A superárvore - que usou as árvores dos genes individuais alinhados - foi construída com o programa CLANN 3.1.3 (Creevey *et al*. 2005) - algoritmo AV - *bootstrap* de 100.

Tabela 3.4 - Metodologias utilizadas para as análises de comparação por similaridade, a obtenção nos genomas de protozoários das sequências dos genes de resistência a drogas e a filogenia

Metodologia	Dados de entrada - Sequências Genes de Resistência a Drogas em Protozoários		Alinhamento	Filogenia
M1	Todos os *hits* obtidos com o hmmpfam 2.3.2		MAFFT 6.240	**PAUP-AV 4.0b10**, *bootstrap* de 10.000
M2	Todos os melhores *hits* obtidos com o hmmpfam 2.3.2	M2-1: **(i)** espécies apresentando o melhor *hit* em todos os genes **(ii)** espécies apresentando o melhor *hit* em quatro genes	(M3)	(M3)
		M2-2: **(A)** sequências de genes completos **(B)** sequências de genes trimados com o GBLOCKS 0.91b **(C)** sequências de genes trimados com o TRIMAL v1.2		
M3	Construção das árvores filogenéticas: individuais, uma supermatriz e uma superárvore	M3-1: Árvores filogenéticas individuais dos grupos **iA, iB, iC, iiA, iiB, iiC**)	MAFFT 6.240	**PAUP-AV 4.0b10**, *bootstrap* de 10.000 **PAUP-MP 4.0b10**, *bootstrap* de 500 **PHYML 2.4.4**, *bootstrap* de 100 **RAXML 7.2.1**, *bootstrap* de 100 **MRBAYES 3.1.2**, n. gerações de 1.000.000 **WEIGHBOR 1.2.1**, *bootstrap* de 10.000
		M3-2: A árvore da supermatriz dos alinhamentos concatenados	MAFFT 6.240	**PAUP-AV 4.0b10**, *bootstrap* de 10.000
		M3-3: A superárvore das árvores dos alinhamentos individuais	MAFFT 6.240	**CLANN 3.1.3**, AV, *bootstrap* de 100

3.2 Filogenômica baseada na reconstrução da árvore de espécies de protozoários

3.2.1 Seleção e preparação dos genes de ortólogos universais e análises filogenéticas

A metodologia para construir a árvore de espécies em protozoários baseou-se no trabalho de Ciccarelli *et al*., (2006) sobre a seleção dos genes universais utilizando o grupo de ortólogos universais COG (*Clusters of Orthologous Groups of proteins*) (anexo 8.13-A, anexo 8.13-B, anexo 8.14). As sequências de proteínas de 31 ortólogos universais envolvidos em processos de tradução foram obtidas da versão mais recente do banco de dados STRING version 8.0 (http://string.embl.de/) e em seguida foram alinhadas com o MAFFT v6.240 (Katoh *et al*. 2005; Katoh *et al*. 2002; Katoh & Toh 2008).

Os perfis HMM (Eddy 1998) foram construídos para os 31 ortólogos universais alinhados e cada um deles foi comparado com os genomas de protozoários disponíveis nos bancos de dados REFSEQ *Release35.catalog* - 01/11/2009 e GENBANK *NCBI-Flat File Release 172.0* (tabela 3.5). Os perfis HMM foram criados e calibrados usando o HMMER versão 2.3.2 - usando os programas hmmbuild e hmmcalibrate- e as comparações foram feitas utilizando o HMMER - usando o programa hmmpfam com *E-value* de 10^{-5}. O anexo 8.13-C mostra a distribuição dos genes ortólogos universais nos genomas de protozoários.

3.2.1.1 Metodologia 1 (M1)

A metodologia M1 apresenta-se esquematizada na tabela 3.6.

Todos os melhores *hits* do hmmpfam dos "31 ortólogos universais em protozoários" foram usados para construir alinhamentos individuais usando o MAFFT v6.240 (Katoh *et al*. 2005; Katoh *et al*. 2002; Katoh & Toh 2008). Os melhores *hits* do hmmpfam referem-se aos primeiros *hits* resultantes da comparação com o HMMER - programa hmmpfam que pertencem, além disso, a um exemplar por espécie de cada um dos táxons dos protozoários encontrados nos 31 ortólogos universais.

Os alinhamentos foram concatenados usando *scripts* nas linguagens PERL e PYTHON. A árvore foi construída com o PHYML 2.4.4 a partir do alinhamento concatenado (Guindon & Gascuel 2003; Guindon *et al*. 2005) com *bootstrap* de 100 e matriz do modelo evolutivo JTT.

3.2.1.2 Metodologia 2 (M2)

Todos os melhores *hits* do hmmpfam dos "31 ortólogos universais em protozoários" foram usados para construir alinhamentos individuais usando MAFFT v6.240 (Katoh *et al*. 2005; Katoh *et al*. 2002; Katoh & Toh 2008). Os melhores *hits* do hmmpfam referem-se aos primeiros *hits* resultantes da comparação com o HMMER - programa hmmpfam que pertencem, além disso, a um exemplar por espécie de cada um dos táxons dos protozoários encontrados nos 31 ortólogos universais.

Os alinhamentos foram trimados com o TRIMAL v1.2 (Capella-Gutierrez *et al*. 2009), visando obter as regiões mais conservadas. Estes, em seguida, foram concatenados usando *scripts* nas linguagens PERL e PYTHON. A árvore foi construída com o PHYML 2.4.4 a partir do alinhamento concatenado (Guindon & Gascuel 2003; Guindon *et al*. 2005), com *bootstrap* de 100 e matriz do modelo evolutivo JTT.

O TRIMAL v1.2 foi usado para eliminar as regiões mais variáveis e/ou divergentes. Todas as posições do alinhamento com *gaps* de 10% ou mais foram removidas, a menos que deixassem menos de 60%. Neste caso, seriam usados apenas os melhores 60%, isto é, com a menor quantidade de *gaps.*

3.2.2 Teste do sinal filogenético

Foram analisados o conteúdo e a distribuição do sinal e o ruído filogenético dos 64 alinhamentos.

Duas abordagens estatísticas - o Teste PTP e a Estatística G1 - foram utilizados para medir o conteúdo total do sinal filogenético. O Teste PTP (*Permutation Test Probability* ou *Permutation Tail Probability Test*) (Faith *et al*. 1991) foi executado com o PAUP 4.0b10 (Swofford 2002), nos parâmetros permute test=ptp, nreps=1000, search=heuristic. A Estatística G1 (Hillis *et al*. 1992) foi calculada a partir dos caracteres usando a função RandTrees do PAUP 4.0b10 com o parâmetro randtrees nreps =1000000.

Tabela 3.5 - Bancos de dados utilizados para a obtenção das sequências dos 31 ortólogos universais dos genomas de protozoários usadas nas análises de comparação por similaridade

Banco de dados	Descrição	Tipo	Número de entradas	Versão	Endereço *web*
STRING	Base de dados de interações de proteínas conhecidas e preditas	Proteína	2.590.259	8.0	http://string.embl.de/
COG	Sistema de classificação baseado nas relações entre os genes ortólogos, Este sistema representa uma análise de genômica comparativa que facilita tanto a anotação funcional de genomas em grande escala, como os estudos evolutivos.	Proteína	4.873	17/04/2003	http://www.ncbi.nlm.nih.gov/COG/
GENBANK	Banco de dados que incorpora sequências de DNA de mais de 105.000 organismos diferentes, principalmente através da submissão direta de dados e as disponibiliza ao público.	Proteína / nucleotídeo	13,5 bilhões bases de nucleotídeos de 12,8 milhões de sequências diferentes	*NCBI-Flat File Release 172.0* - 06/15/2009	http://www.ncbi.nlm.nih.gov/Genbank/
KOG	Versão do COG, para sete genomas eucarióticos completos: *S. cerevisiae*, *S, pombe*, *E. cuniculi*, *A. thaliana*, *C. elegans*, *D. melanogaster* e *H. sapiens*.	Proteína	4.852	21/07/2003	http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi
REFSEQ	Banco de dados de proteínas não redundantes do NCBI	Proteína	9.662.677	*Release35.catalog* - 01/11/2009	http://www.ncbi.nlm.nih.gov/RefSeq/

Tabela 3.6 - Metodologias utilizadas para as análises de comparação por similaridade e a obtenção das sequências dos 31 ortólogos universais

Metodologia	Dados de entrada - Sequências "31 ortólogos universais em protozoários"	Alinhamento	Filogenia
M1	Todos os melhores *hits* do hmmpfam dos 31 ortólogos universais totais concatenados	MAFFT v6.240	**PHYML 2.4.4**, *bootstrap* de 100
M2	Todos os melhores *hits* do hmmpfam dos 31 ortólogos universais trimados e concatenados	MAFFT v6.240	**PHYML 2.4.4**, *bootstrap* de 100

3.3 Filogenômica dos EGM em protozoários

3.3.1 Seleção e preparação dos genes encontrados nos EGM em protozoários e análises filogenéticas

Foram usadas as sequências de aminoácidos dos genes EGM de todos os organismos disponíveis. Estas sequências foram obtidas dos bancos de dados SWISS-PROT e TREMBL - release11.0 - (http://www.expasy.ch/sprot/) e pertencem aos seguintes genes: (1) transcriptase reversa, (2) telomerase, (3) ribonuclease H, (4) integrase, (5) proteína gag, (6) proteína gag-pol e (7) proteína pol.

Para cada grupo de genes EGM, foram criados e calibrados perfis HMM (Eddy 1998) usando o programa HMMER versão 2.3.2 - hmmbuild e hmmcalibrate. Cada um destes perfis HMM foi comparado aos genomas de protozoários catalogados e disponíveis nos bancos de dados REFSEQ e GENBANK e nos arquivos taxdump do banco de dados *Taxonomy* (ftp://ftp.ncbi.nih.gov/pub/taxonomy/) (tabela 3.7), utilizando o HMMER - programa hmmpfam com *E-value* de 10^{-5}.

3.3.1.1 Metodologia

Todos os *hits* do hmmpfam dos Genes de EGM em Protozoários foram usados para construir alinhamentos individuais usando o MAFFT v6.240(Katoh *et al*. 2005; Katoh *et al*. 2002; Katoh & Toh 2008). Todos os *hits* do hmmpfam referem-se a todos os *hits* resultantes da comparação com o HMMER - programa hmmpfam que pertencem a todos os táxons dos protozoários encontrados nos genes pertencentes aos EGM.

O ARPA foi usado para construir árvores filogenéticas usando os seguintes programas:

- PAUP-AV 4.0b10 com *bootstrap* de 10.000,
- PAUP-MP 4.0b10 com *bootstrap* de 500,
- PHYML 2.4.4 com *bootstrap* de 100,
- RAXML 7.2.1 com *bootstrap* de 100,
- MRBAYES 3.1.2 com número de gerações de 10.000 e
- O fluxograma de execução do WEIGHBOR com *bootstrap* de 10.000 (que inclui o programa SEQBOOT do pacote PHYLIP 3.66 (Felsenstein 1989), os programas TREE-PUZZLE

5.2 (Schmidt *et al*. 2002; Schmidt & von Haeseler 2007) e WEIGHBOR 1.2.1 (Bruno *et al*. 2000) e o programa CONSENSE do pacote PHYLIP 3.66).

As matrizes de modelo evolutivo e outros parâmetros evolutivos foram obtidos com o programa MODELGENERATOR version 0.85. Estes foram usados nos programas PHYML 2.4.4, RAXML 7.2.1 e MRBAYES 3.1.2 e no fluxograma de execução do WEIGHBOR, nos seguintes genes EGM:

- transcriptase reversa, matriz WAG e distribuição gama,

- proteína gag, matriz WAG e modelo de heterogeneidade de taxa de sítios invariáveis,

- proteína gag-pol, matriz WAG, distribuição gama e modelo de heterogeneidade de taxa de sítios invariáveis,

- proteína pol, matriz BLOSUM62 e distribuição gama,

- integrase, matriz BLOSUM62 e distribuição gama e

- ribonuclease H, matriz BLOSUM62, distribuição gama e modelo de heterogeneidade de taxa de sítios invariáveis.

As árvores foram editadas utilizando os programas MEGA 4 (Kumar *et al*. 2008) e TREEVIEW X 0.5.0 (Kumar *et al*. 2008) (Page 1996; Page 2002).

Tabela 3.7 - Bancos de dados utilizados para a obtenção das sequências dos genes de resistência a drogas e os genomas de protozoários usadas nas análises de comparação por similaridade

Banco de dados	Descrição	Tipo	Número de entradas	Versão	Endereço *web*
SWISS-PROT e TREMBL	Banco de dados de proteínas curadas	Proteína	181.163.771	release11.0	http://www.expasy.ch/sprot/
TAXONOMY	Banco de dados do NCBI que contém os nomes de todos os organismos que estão representados nas bases de dados genéticos com pelo menos uma sequência de nucleotídeos ou proteínas.	---	---	06/13/2006	http://www.ncbi.nlm.nih.gov/taxonomy
REFSEQ	Banco de dados de proteínas não redundantes do NCBI	Proteína	9.662.677	*Release35.catalog* - 01/11/2009	http://www.ncbi.nlm.nih.gov/RefSeq/
GENBANK	Banco de dados que incorpora sequências de DNA de mais de 105.000 organismos diferentes, principalmente através da submissão direta de dados e as disponibiliza ao público.	Proteína / nucleotídeo	13,5 bilhões bases de nucleotídeos de 12,8 milhões de sequências diferentes	*NCBI-Flat File Release 172.0* - 06/15/2009	http://www.ncbi.nlm.nih.gov/Genbank/

3.3.2 Estratégia para a detecção da seleção positiva de genes dos EGM em protozoários

3.3.2.1 Grupo de dados biológicos e análise de pressão seletiva

Foi utilizado o método da máxima verossimilhança para examinar as forças seletivas que atuam nos seguintes genes de EGM encontrados em protozoários: (1) telomerase, (2) transcriptase reversa total, (3) transcriptase reversa não-LTR, (4) transcriptase reversa LTR, (5) proteína gag, (6) proteína gag-pol, (7) integrase e (8) ribonuclease H. O programa usado foi o CODEML do pacote PAML v.4 (Yang 2007). Para estimar a taxa ω em todos os sítios de cada árvore filogenética foram usados os modelos de evolução de sequências M0 (razão única), M1 (aproximadamente neutro), M2 (seleção positiva), M3 (discreto), M7 (beta) e M8 (beta&ω) (Yang *et al*. 2000).

As sequências que apesentaram *stop codons* foram excluídas da análise com o programa transeq (http://www.ebi.ac.uk/Tools/emboss/transeq/index.html), para serem aceitas e executadas pelo CODEML (tabela 3.8). A proteína pol e o íntron do grupo II foram excluídas da análise devido a que todas as sequências apresentaram *stop codons*.

Duas vezes o valor de *log-likelihood* das comparações entre os modelos M0:M3, M1:M2 e M7: M8 foram avaliadas por significância estatística para determinar qual delas se ajusta melhor aos dados e se as sequências exibiram seleção positiva, respectivamente. A razão estatística de *log-likelihood* são assumidos como $\chi 2$ distribuído com graus de liberdade igual à diferença no número de parâmetros entre os modelos. Foi usada comparação M0 (razão única) com o M3 (discreta) para testar a pressão seletiva variável entre os sítios e três LRT para testar sítios em evolução por seleção positiva, comparando-se (i) M1 (neutro) contra M2 (seleção positiva) e (ii) M7 (beta) contra M8 (beta&ω). Estimativas de máxima verossimilhança dos parâmetros e as contagens de probabilidade para os genes são apresentados nas tabelas 4.7-4.14.

Tabela 3.8 - Sequências usadas na análise para a detecção da seleção positiva dos genes de EGM

Gene	Sequências usadas na análise * (sem *stop codom*)	Tamanho do alinhamento das sequências	Sequências excluídas da análise (com *stop codom*)
Telomerase	**25** Cpa (2), Cho (2), Gla (3), Lam (1), Ldo (1), Lbr (3), Lma (3), Lin (1), Tan (2), Tcr (3), Tbr (3)	4.833	0
Transcriptase reversa total	**122** Cpa (2), Cho (2), Ete (2), Ehi (2), Gla (24), Han (1), Lam (1), Ldo (1), Lbr (5), Lma (3), Lin (1), Ppa (3), Pin (13), Pra (3), Pso (3), Pyo (2), Pye (2), Pli (18), Rsp (1), Rsa (9), Tth (1), Tan (2), Tva (2), Tcr (8), Tbr (10)	6.657	37
Transcriptase reversa não-LTR	**45** Ehi (2), Gla (21), Lbr (2), Pin (2), Pyo (2), Tth (1), Tva (2), Tcr (5), Tbr (8)	7.479	0
Transcriptase reversa LTR	**30** Ete (2), Ppa (3), Pin (11), Pra (3), Pso (3), Pye (2), Rsa (5)	5.346	0
Proteína gag	**22** Ddi (4), Ptr (8), Pin (2), Tps (7)	5.349	0
Proteína gag-pol	**14** Ptr (11), Tps (2)	6.780	0
Integrase	**27** Pyo (1), Tva (24), Pyo (1)	3.828	21
Ribonuclease H	**37** Bbo (2), Gla (4), Lin (2), Lma (8), Lbr (3), Pyo (2), Tva (2), Tbr (6), Tcr (7)	2.952	14

*Incluída uma sequência *output* usada para enraizar a árvore

3.3.2.2 Estratégia para a detecção *in silico* da seleção positiva de genes dos EGM em protozoários

Uma das principais etapas na análise das sequências de qualquer gene em estudo é determinar que tipo de pressão seletiva está sendo exercida em sua sequência e, portanto, prever seu comportamento evolutivo. Na análise dos dados de sequência, a melhor forma de detectar a ação da seleção natural é através de casos documentados em que haja um maior número de substituições não-sinônimas (envolvendo uma mudança de aminoácidos) (d_N) em relação ao número de substituições sinônimas (d_S) dentro do gene analisado. Isto sugere a presença de seleção positiva, pois porque significa que a taxa de substituição é maior do que a taxa de mutações, o que não poderia ocorrer se as mutações fossem neutras (onde as taxas são iguais). No caso da seleção negativa ou purificadora, fixa-se um maior número de substituições d_S do que de d_N, porque estas últimas são geralmente deletérias e, portanto, seletivamente removidas.

3.3.2.3 Métodos de detecção de seleção

A comparação da razão entre o número de substituições sinônimas (silenciosas) e não-sinônimas (com a alteração de aminoácidos) é considerada um dos métodos mais sensíveis de detecção de seleção natural usado no estudo dos mecanismos da evolução das sequências de DNA (Gillespie 1992; Kimura 1989; Ohta 1993). As mutações sinônimas não são detetadas pela seleção natural (Akashi 1995; Akashi 1994), enquanto as mutações não-sinônimas podem estar sob uma forte pressão seletiva. A comparação da fixação das razões desses dois tipos de mutações fornece uma ferramenta poderosa para entender o efeito da seleção natural na evolução molecular das sequências.

Uma medida que tem sido usada de forma contínua e intensa em vários estúdios é a taxa de substituição não-sinônima/sinônima ($\omega = d_N / d_S$) (Miyata *et al*. 1980). Aqui as taxas d_N e d_S são definidos como os números de substituições não-sinônimas e sinônimas por sítio, e a sua razão ω mede a pressão seletiva no nível de proteína. Um $\omega > 1$ significa que as mutações não sinônimas oferecem vantagens para as proteínas e possuem uma maior probabilidade de fixação do que as mutações sinônimas (Yang *et al*. 2000). O regime de seleção é determinado pela razão entre d_N / d_S (ω), sendo:

(i) $\omega > 1$ para seleção positiva.

(ii) $\omega < 1$ para seleção negativa ou purificadora.

(iii) $\omega = 1$ para um regime neutro de substituições.

3.3.2.4 Determinação de d_N e d_S por verossimilhança

A estimativa de d_N *e* d_S por verossimilhança (Goldman *et al*. 1994) baseia-se em um modelo explícito de substituição de códons (matriz de substituição). Entretanto, os modelos usados para estimar d_N *e* d_S partem da premissa de que todos os sítios estão sujeitos à mesma pressão de seleção, o que é pouco realista, visto que regiões de uma mesma proteína desempenham funções distintas e, com isso, a pressão evolutiva é diferente ao longo de uma sequência protéica. O uso destes modelos na detecção de seleção positiva é pouco sensível e geralmente não detecta seleção positiva quando esta atua em alguns sítios de uma proteína (Nielsen 1997a; Nielsen 1997b). Nielsen e Yang (1998) propuseram um método onde os sítios (códons) de uma proteína podem assumir valores de d_N *e* d_S distintos. Com isso, a detecção de seleção positiva pode ser feita individualmente para cada um dos sítios (códons) de uma determinada proteína. Este método está implementado no programa CODEML, do pacote PAML (Yang 2007). O CODEML utiliza vários modelos evolutivos que diferem entre si na distribuição de d_N/d_S (ω) entre os códons. Além disso, é possível a incorporação das diferenças nas taxas de heterogeneidade entre os sítios pela estimativa do parâmetro alfa da distribuição gama. O método de verossimilhança usa a correção filogenética para descrever o processo de evolução dos códons, proporcionando desta forma maior confiabilidade da hipótese formulada (Freckleton *et al*. 2002). Os vários modelos de substituição de códons implementados no programa CODEML são (Nielsen 1997a; Nielsen 1997b; Yang *et al*. 2000):

(i) M0: é o mais simples, pois assume um único valor de d_N/d_S para todos os sítios (códons).

(ii) M1 (Neutro): assume duas categorias de códons - os conservados ($d_N/d_S = 0$) e os neutros ($d_N/d_S = 1$), distribuídos em diferentes proporções (p_0 e p_1).

(iii) M2 (seleção): comparado ao modelo anterior, adiciona uma nova categoria de códons (p_2) que pode assumir valores maiores que 1 para d_N/d_S (ω_2).

(iv) M3 (discreto): permite n categorias de códons para um determinado número de classes (K), os valores de d_N/d_S (ω) por categoria, são estimados a partir dos dados e seguem uma distribuição gama, subdividida em K categorias.

(v) M7: assume dez categorias de códons, com valores de d_N/d_S distribuídos nestas categorias de acordo com uma distribuição beta. Além disso todos os valores são menores que 1. De modo que este modelo indica seleção negativa.

(vi) M8: incorpora uma décima-primeira categoria a mais que o modelo M7, a qual pode ter valores de d_N/d_S maiores que 1. Permite, assim, a detecção de seleção positiva na última categoria.

A abordagem de detecção de seleção positiva é inicialmente feita pela estimativa de parâmetros para todos os modelos acima citados. O teste LRT (*Likelihood Ratio Test*), usado para verificar qual o modelo que melhor se ajusta aos dados, é feito pela diferença dos valores de verossimilhança entre os modelos aninhados. Visto que essa diferença segue uma distribuição chi-quadrado, permite determinar a significância estatística entre os modelos, assumindo que os graus de liberdade são dados pela diferença entre os parâmetros de cada modelo. Os modelos são testados da seguinte forma e com os seguintes graus de liberdade:

(i) M0 vs M2 com 2 graus de liberdade

(ii) M0 vs M3 com 4 graus de liberdade

(iii) M1 vs M2 com 2 graus de liberdade

(iv) M1 vs M3 com 4 graus de liberdade

(v) M2 vs M3 com 3 graus de liberdade

(vi) M7 vs M8 com 2 graus de liberdade

3.3.2.5 *Likelihood Ratio Test* (LRT)

O "melhor" modelo evolutivo é indicado pelo teste de razão de verossimilhança LRT (*Likelihood Ratio Test*). Este teste é usado nas inferências estatísticas e deve ser aplicado somente quando os modelos são aninhados. Dois modelos são aninhados quando o mais complexo (aquele com um número maior de parâmetros livres) deriva de um mais simples (com menos parâmetros livres) e quando no mais complexo existe a adição de pelo menos um parâmetro. Após a estimativa dos valores de verossimilhança do modelo mais simples (L0) e do mais complexo (L1), estes são tratados no teste LRT como hipótese nula (H_0) e hipótese alternativa (H_1), respectivamente. Isto acontece devido à diferença de valores do logaritmo de verossimilhança entre dois modelos aninhados na seguinte equação:

$$\mathbf{LRT=(2[lnL1-lnL0])}$$

Os valores de LRT seguem uma distribuição qui-quadrado (ω_2), onde os graus de liberdade são dados pela diferença de parâmetros entre os modelos. A hipótese nula é descartada quando o valor do teste for menor que o nível de significância escolhido (valor de alfa).

A detecção de seleção positiva ocorre se o modelo escolhido assume parte da premissa de que uma determinada proporção de sítios tem valores de d_N / d_S maiores que 1 (ω>1). Assim, se o teste LRT indicar que o modelo M2 é mais adequado do que os modelos M0 e M1, então a hipótese de seleção positiva é aceita. Assim, além da indicação de seleção positiva teremos a indicação de quais os códons que têm valores de d_N/d_S significativamente acima de um. A melhor comparação para determinar a presença de seleção positiva em um alinhamento é feita entre M0 vs. M2 e M1 vs. M2, pois o modelo M2 é mais conservador na estimativa de valores de d_N/d_S (Yang *et al*. 2000). De sorte que, somente após isso, poderemos inferir com mais confiabilidade os códons sob seleção (M3 e M8). O método de Yang possibilita a adequação de modelos de substituição de códons (uso de diferentes matrizes de transição de aminoácidos) e o ajuste das taxas de heterogeneidade nas estimativas de ω para cada códon, bem como o uso da informação genealógica do alinhamento para descrever o processo evolutivo de substituição de códons (Yang 1994a; Yang 1994b). Contudo, um dos aspectos negativos do método de verossimilhança na detecção de seleção positiva é a possibilidade de inflação nos valores de d_N, especialmente quando o número de amostras é baixo (>20) e quando não existe sinal filogenético.

3.4 A telomerase e os elementos de retrotransposição em Tri-tryps baseados em análises filogenéticas dos domínios da transcriptase reversa

3.4.1 Seleção e preparação dos genes encontrados nos EGM nos Tri-tryps e análises filogenéticas

As sequências de dados de proteínas pertencentes à transcriptase reversa foram obtidas dos bancos de dados GENBANK (http://www.ncbi.nlm.nih.gov/) e PFAM (Bateman *et al*. 2004) e baseadas em trabalhos publicados por Eickbush e Malik (Malik *et al*. 1999; Xiong *et al*. 1990).

3.4.1.1 Metodologia

O alinhamento da família PF00078-RVT_1 - *Reverse transcriptase (RNA-dependent DNA polymerase)* formado pelos oito domínios característicos da transcriptase reversa foi obtido do banco de dados PFAM 23.0 (http://pfam.sanger.ac.uk/).

As sequências de aminoácidos dos seguintes genes encontrados nos EGM em Tri-tryps foram obtidas dos bancos de dados SWISS-PROT e TREMBL - release11.0 - (http://www.expasy.ch/sprot/):

(i) transcriptase reversa, separada em dois grupos (Xiong & Eickbush 1990):

(ia) não-LTR e

(ib) LTR e

(ii) telomerase.

Foram construídos alinhamentos individuais para os grupos **ia**, **ib** e **ii** com o MAFFT v5.861 (Katoh *et al*. 2005; Katoh *et al*. 2002; Katoh & Toh 2008). Estes três alinhamentos foram adicionados ao alinhamento PF00078-RVT_1, usando o algoritmo perfil-perfil do programa CLUSTALW 1.83 (Thompson *et al*. 2002; Thompson *et al*. 1994).

Visando uma análise mais detalhada da posição dos Tri-tryps, os 227 táxons pertencentes ao alinhamento constituído pelos grupos **ia**, **ib**, **ii** e PF00078-RVT_1 foram reduzidos para 84 táxons internos (*ingroup*) e a um conjunto de táxons externos (*outgroup*) formados pelos eucariotos: *Arabidopsis*, *Drosophila*, *Escherichia coli* e *Synechocystis*.

Foram usadas todas as espécies de tripanossomatídeos e as sequências das espécies representativas dos eucariotos, a fim de evitar duplicações e minimizar os artefatos de atração dos ramos longos (*Long Branch Attraction*).

3.4.1.2 Análises filogenéticas

As árvores filogenéticas foram construídas usando os programas:

- PAUP-AV 4.0b10 com 10.000 de *bootstrap* (Saitou *et al*. 1987),

- PAUP-MP 4.0b10 com 500 de *bootstrap*

- PHYML 2.4.4 com 100 de *bootstrap* (Guindon & Gascuel 2003; Guindon *et al*. 2005),

- MRBAYES 3.1.2 com 1,000.000 de número de gerações (Ronquist & Huelsenbeck 2003). A matriz WAG foi usada na execução dos algoritmos de MV e IB. O gene hsp70 foi a exceção devido a que foi usada a matriz RtREV e

- o fluxograma de execução do WEIGHBOR com 10.000 de *bootstrap* (Bruno *et al*. 2000; Felsenstein 1989; Schmidt *et al*. 2002; Schmidt & von Haeseler 2007),

A matriz de distância evolutiva foi construída usando o modelo BLOSUM62 (Henikoff *et al*. 1992), obtido dos critérios AIC (*Akaike Information Criterion*) e *Hierarchical Likelihood Ratio Tests*, os quais foram selecionados utilizando o programa MODELGENERATOR version 0.85 (Keane *et al*. 2006).

A inferência filogenética foi baseada nas topologias e valores de consistência das cinco árvores filogenéticas.

(i) AV usando PAUP

A árvore de AV foi construída com o modelo BLOSUM62 e o programa PAUP 4.0b10. As análises foram realizadas executando os seguintes parâmetros: permutação dos braços TBR (*Tree-Bisection-Connection*), *gaps* tratados como perdidos (*missing*) e *bootstrap* de 10.000. Foi feita uma busca heurística. A árvore consenso foi obtida usando 50% *Majority-rule consensus* com o programa contree do pacote PHYLIP 3.66.

(ii) MP usando PAUP

A árvore de MP foi construída usando o PAUP 4.0b10. As análises foram realizadas usando os seguintes parâmetros: permutação dos braços TBR (*Tree-Bisection-Connection*), *gaps* tratados como perdidos (*missing*) e *bootstrap* de 500. Foi feita uma busca heurística. A árvore

consenso foi obtida usando 50% *Majority-rule consensus* com o programa contree do pacote PHYLIP 3.66.

(iii) MV usando PHYML

A árvore de MV foi construída com o programa PHYML 2.4.4, usando os seguintes parâmetros: *bootstrap* de 100, modelo evolutivo BLOSUM62, distribuição gama com quatro categorias e sítios invariáveis e parâmetro alfa da função gama.

(iv) Fluxograma de Execução do WEIGHBOR

Foi executado o SEQBOOT com um *bootstrap* de 10.000. As distâncias respectivas foram calculadas com o TREE-PUZZLE 5.2 (Schmidt *et al*. 2002; Schmidt & von Haeseler 2007), usando os seguintes parâmetros: distribuição gama discreta de seis categorias mais sítios invariáveis, parâmetro alfa da função gama, fração de sítios invariáveis, taxas relativas das categorias e atribuição de categoria para cada sítio. Além disso, para estimar valores de suporte de cada nó, foi utilizado o método QP (*Quartet-Puzzling*). A análise de distância foi feita com o programa WEIGHBOR 1.2.1. A árvore-consenso foi obtida usando 50% *Majority-rule consensus* com o programa contree do pacote PHYLIP 3.66.

(v) IB usando MRBAYES

A Inferência Bayesiana foi realizada utilizando o programa MRBAYES 3.1.2, com parâmetros definidos para quatro cadeias, com número de gerações de 1.000.000, matriz BLOSUM62, distribuição gama, proporção de sítios invariáveis, valores de *burn-in* de 1.000 gerações e árvores salvas a cada 100 gerações. Para a análise de inferência Bayesiana foram usados os critérios AIC (*Akaike Information Criterion*) e *Hierarchical Likelihood Ratio Tests*.

3.5 ProtozoaDB: A visualização dinâmica e exploração dos genomas dos protozoários

O esquema Phylo foi projetado pelo nosso grupo de trabalho para armazenar todos os dados relacionados a um estudo filogenético.

Foram estudados os dados pertencentes a todas as colunas e tabelas do esquema Phylo, assim como as respectivas relações destes dados com as informações encontradas nas árvores filogenéticas.

Aquelas tabelas que armazenam as informações obtidas diretamente das árvores filogenéticas: tree, tree-edge, tree-taxa e tree-analysis foram o centro deste estudo.

Para extrair e formatar as informações contidas nas árvores filogenéticas foram feitos *scripts* nas linguagens PERL e PYTHON e usadas as bibliotecas BIOPERL e BIOPYTHON.

CAPÍTULO 4 - RESULTADOS

4.1 Reconstrução automática de análises filogenômicas (ARPA) de genes relacionados à resistência a drogas em genomas de protozoários

4.1.1 Projeto ARPA

O ARPA é um sistema para a análise filogenética de dados moleculares que integra diferentes algoritmos filogenéticos executados por programas. As abordagens filogenéticas mais representativas incluídas na execução do ARPA são os algoritmos filogenéticos, otimização de comprimento dos ramos, testes do modelo evolutivo e métodos de reconstrução filogenética baseados em distância, máxima parcimônia, máxima verossimilhança e métodos bayesianos. Os programas incluídos no ARPA para a reconstrução filogenética são: GARLI, MRBAYES, PAUP, PHYLIP, PHYML, RAXML, TREE-PUZZLE e WEIGHBOR.

O ARPA pode ser utilizado para análises filogenéticas de sequências de nucleotídeo ou aminoácido usando um ou vários genes concatenados, e foi desenvolvido para ser executado: (i) localmente, através de linhas de comando e (ii) por meio da interface *web*. A estrutura de linha de comando e os algoritmos executados pela interface *web* foram desenvolvidos usando a linguagem de programação PYTHON.

4.1.2 Metodologia 1 (M1)

Foram obtidas as árvores filogenéticas AV de todos os *hits* do hmmpfam dos seguintes Genes de Resistência a Drogas em Protozoários: AQP, hsp70, GP63, TRYR e MRPA. Os táxons - espécies de protozoários - pertencentes às árvores filogenéticas dos genes hsp70 e MRPA (figura 4.1, anexo 8.3) foram muito numerosos, pelo qual foi necessário separá-las em grupos para uma melhor análise (anexo 8.6). A tabela 4.1 mostra a quantidade de táxons da árvore assim como a quantidade de táxons dos grupos, que estão em parênteses. O grupo rotulado como **Vários** refere-se aos grupos de diferentes espécies considerados como grupos não resolvidos (anexo 8.3, 8.6; tabela 4.2, 4.3).

Tabela 4.1 - Genes de resistência a drogas usados para a filogenia

Gene	N. táxons	Táxons
MRPA	942	Tabela 4.2
AQP	104	*D. discoideum* (9), *Entamoeba* (4), *Leishmania* (30): [*L. major*, *L. infantum*, *L. brasiliensis*], *M. brevicollis* (3), *Paramecium* (25), *Plasmodium* (13), *T. thermophila* (1), *T. gondii* (3) e *Trypanosoma* (16): [*T. cruzi*, *T. brucei*],
GP63	480	*D. discoideum* (7), *Entamoeba* (3), *Leishmania* (183): [*L. brasiliensis* (84), *L. donovani* (39), *L. infantum* (32), *L. major* (9), *L. guyanensis* (5), *L. donovani chagasi* (5), *L. panamensis* (3), *L. tropica* (2), *L. sp* (2), *L. arabica* (1), *L. aethiopica* (1), *L. mexicana* (1)], *P. tetraurelia* (13), *T. thermophila* (56), *T. vaginalis* (56) e *Trypanosoma* (161): [*T. cruzi* (143), *T. brucei* (18)]
hsp70	855	Tabela 4.3
TRYR	256	*Babesia* (3), *Cryptosporidium* (2): [*C. hominis*, *C. parvum*], *D. discoideum* (9), *Entamoeba* (12): [*E. dispar* (4), *E. histolytica* (8)], *Giardia* (6), *Leishmania* (33): [*L. brasiliensis* (12), *L. major* (9), *L. infantum* (8), *L. amazonensis* (2), *L. donovani* (2)], *Monosiga* (9), *P. tetraurelia* (13), *Plasmodium* (48): [*P. falciparum* (18), *P. berghei* (9), *P. vivax* (5), *P. knowlesi* (5), *P. chaubadi* (6), *P. yoelii* (5)], *Theileria* (6): [*T. annulata* (3), *T. parva* (3)], *T. thermophila* (9), *Toxoplasma* (4), *T. vaginalis* (12) e *Trypanosoma* (87): [*T. cruzi* (77), *T. brucei* (12) e *T. congolensi* (1)]

4.1.2.1 Transportadores MRPA

A árvore do gene MRPA (942 espécies) apresentou 36 grupos melhor definidos (tabela 4.2, figura 4.1, anexo 8.3).

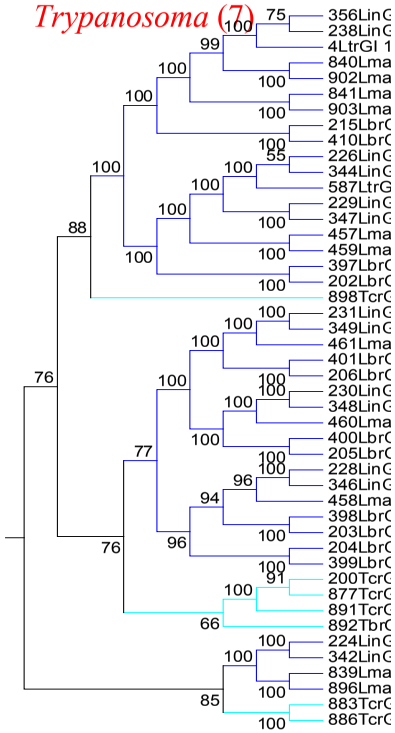
Tabela 4.2 - Os 36 grupos da árvore filogenética AV do gene MRPA

Número do Grupo	Nome do Grupo: Gênero / Espécie	Número de espécies nos gêneros	Espécies	Observações do grupo
4	*Leishmania* *Trypanosoma*	38 7	*L. infantum*, *L. major*, *L. brasiliensis*, *L. tropica* e *T. cruzi*	Formou dois subgrupos formados pelas cinco espécies
8	*Leishmania* *Trypanosoma*	5 3	*L. infantum*, *L. major*, *L. brasiliensis*, *L. tropica* e *T. cruzi*	Evento de duplicação separou *Leishmania* de *Trypanosoma*
32	*Leishmania* *Trypanosoma*	40 9	*L. infantum*, *L. major*, *L. brasiliensis*, *L. tropica*, *L. amazonensis*, *L. tarentolae*, *L. guyanensis*, *T. cruzi* e *T. brucei*	Formou três subgrupos formados pelas espécies mencionadas
9	*Leishmania*	3	*L. major* e *L. brasiliensis*	Formado pelas espécies mencionadas
29	*Leishmania*	7	*L. infantum*, *L. major* e *L. brasiliensis*,	Apresentou monofilia
1	**Vários** *Leishmania* *Trypanosoma*	**36** 10 7	*L. infantum*, *L. major*, *L. brasiliensis*, *T. cruzi* e *T. brucei*	Formou três subgrupos formados pelas espécies mencionadas
10	**Vários** *Leishmania* *Trypanosoma* *D. discoideum*	11 11 24	*L. infantum*, *L. major*, *L. brasiliensis*, *T. cruzi*, *T. brucei* e *D. discoideum*	Os dois gêneros e a espécie formaram subgrupos monofiléticos
17	**Vários** *Leishmania* *Trypanosoma* *Plasmodium*	**33** 5 3 10	*L. infantum*, *L. major*, *L. brasiliensis*, *T. cruzi*, *T. brucei*, *P. falciparum*, *P. vivax*, *P. knowlesi*, *P. chabaudi* e *P. yoelii*	Os três gêneros formaram subgrupos monofiléticos
15	**Vários** *T. vaginalis* *D. discoideum*	12 14	*T. vaginalis* e *D. discoideum*	As duas espécies formaram subgrupos monofiléticos
18	**Vários** *D. discoideum* *Theileria*	6 3	*D. discoideum*, *T. annulata* e *T. parva*	As três espécies formaram subgrupos monofiléticos
25	**Vários**		*C. parvum*, *C. hominis*, *T. annulata*, *P. falciparum*, *P. vivax*, *P. knowlesi*, *P.*	Os três gêneros formaram

	Cryptosporidium *Theileria* *Plasmodium*	18 2 11	*berghei* e *P. yoelii*	subgrupos monofiléticos
27	**Vários** *Toxoplasma* *Entamoeba* *Leishmania* *Dictyostelium*	 8 24 19 11	*T. gondii*, *E. dispar, E. histolytica*, *L. infantum*, *L. major*, *L. brasiliensis*, *L. donovani*, *L. amazonensis*, *L. enriettii*, e *D. discoideum*	Os quatro gêneros formaram subgrupos monofiléticos
2	*P. tetraurelia* *T. thermophila*	30 54	*P. tetraurelia* e *T. thermophila*	As duas espécies formaram subgrupos monofiléticos
5	*T. thermophila* *P. tetraurelia* *B. brevicollis*	18 4 3	*T. thermophila*, *P. tetraurelia* e *B. brevicollis*	As três espécies formaram subgrupos monofiléticos
7	*T. thermophila* *P. tetraurelia*	10 17	*T. thermophila* e *P. tetraurelia*	As duas espécies formaram subgrupos monofiléticos
11	*Plasmodium* *Tetrahymena* *Paramecium*	31 12 3	*P. falciparum*, *P. vivax*, *P. knowlesi*, *P. berghei*, *P. chabaudi*, *T. thermophila* e *P. tetraurelia*	Os três gêneros formaram subgrupos monofiléticos
13	*Plasmodium* *Tetrahymena*	11 4	*P. falciparum*, *P. vivax*, *P. knowlesi*, *P. berghei*, *P. chabaudi*, *P. yoelii* e *T. thermophila*	Os dois gêneros formaram subgrupos monofiléticos
14	*Tetrahymena* *Paramecium* *T. gondii*	11 7 1	*T. thermophila*, *P. tetraurelia* e *T. gondii*	As três espécies formaram subgrupos monofiléticos
6	*D. discoideum*	12	*D. discoideum*	Formou monofilia
37	*D. discoideum*	9	*D. discoideum*	Formou monofilia
34	*D. discoideum* *P. tetraurelia* *E. histolytica* *T. thermophila*	16 15 2 6	*D. discoideum*, *P. tetraurelia*, *E. histolytica* e *T. thermophila*	As quatro espécies formaram subgrupos monofiléticos
16	*T. vaginalis*	17	*T. vaginalis*	Formou monofilia
20	*T. vaginalis*	16	*T. vaginalis*	Formou monofilia
22	*Plasmodium*	5	*P. falciparum*, *P. vivax* e *P. knowlesi*	Formou monofilia

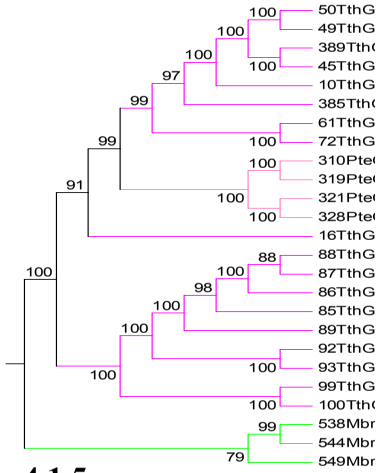
26	*Plasmodium*	6	*P. falciparum*, *P. vivax*, *P. berghei*, *P. yoelii* e *P. knowlesi*	Formou monofilia
21	*Plasmodium* *E. dispar*	14 1	*P. falciparum*, *P. vivax*, *P. knowlesi*, *P. berghei*, *P. chabaudi*, *P. yoelii* e *E. dispar*	Formou monofilia
23	*B. bovis*	3	*B. bovis*	Formou monofilia
24	*Theileria*	30	*T. annulata* e *T. parva*	Formou monofilia
28	*Cryptosporidium*	4	*C. parvum* e *C. hominis*	Formou monofilia
35	*Cryptosporidium*	4	*C. parvum* e *C. hominis*	Formou monofilia
31	*G. lamblia*	4	*G. lamblia*	Formou monofilia
30	*G. lamblia* *D. discoideum*	5 3	*G. lamblia* e *D. discoideum*	As duas espécies formaram subgrupos monofiléticos
33	*Entamoeba* *T. vaginalis*	14 3	*E. histolytica*, *E. dispar* e *T. vaginalis*	O gênero e a espécie formaram subgrupos monofiléticos
36	*M. brevicollis*	5	*M. brevicollis*	Grupo monofilético
3	**Vários**	**10**	*T. vaginalis*, *C. parvum*, *C. hominis*, *G. lamblia*, *C. merolae* e *C. paradoxa*	Não formaram subgrupos determinados
12	**Vários**	**38**	*P. tetraurelia, T. thermophila, C. hominis, C. parvum, L. brasiliensis, T. cruzi, T. brucei, D. discoideum, E. dispar, E. histolytica, G. lamblia, T. gondii* e *M. brevicollis*	As espécies formaram pequenos subgrupos monofiléticos
19	**Vários**	**14**	*C. parvum*, *C. hominis*, *E. histolytica*, , *T. cruzi, T. brucei, M. brevicollis*, *D. discoideum*, *T. thermophila, T. gondii* e *P. falciparum*	Não formaram subgrupos determinados

4.- *Leishmania* (38), *Trypanosoma* (7)



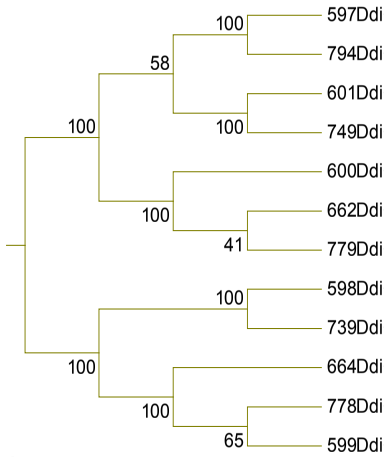
4.1.4

5.- *T. thermophila* (18),



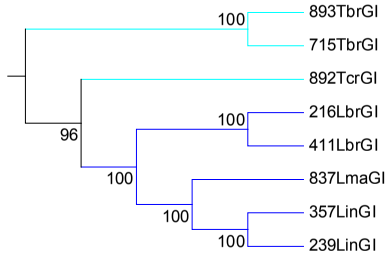
4.1.5

6.- *D. discoideum* (12) *P. tetraurelia* (4),*M. brevicolis* (3)



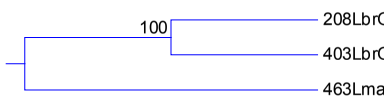
4.1.6

8.-*Leishmania*(5), *Trypanosoma*(3)



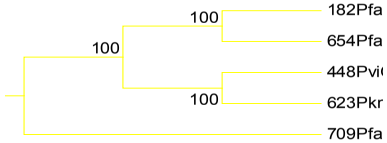
4.1.8

9.- *Leishmania* (3)



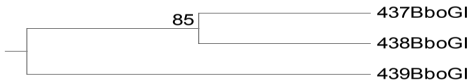
4.1.9

22.- *Plasmodium* (5)



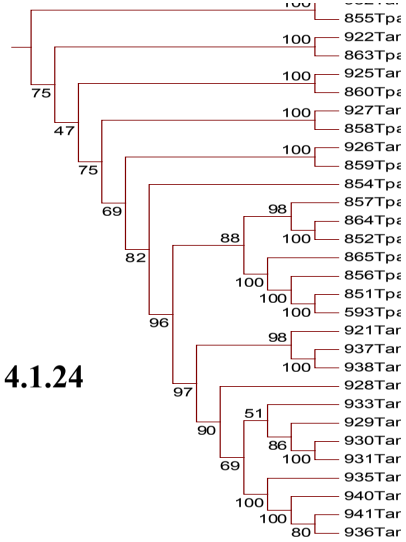
4.1.22

23.- *B. bovis* (3)



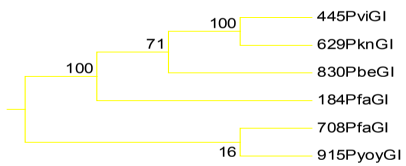
4.1.23

24.- *Theileria* (30)



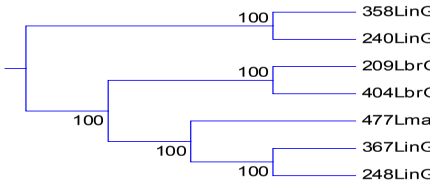
4.1.24

26.- *Plasmodium* (6)



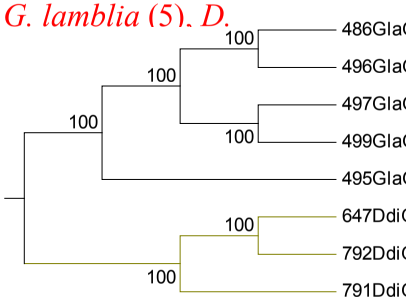
4.1.26

29.- *Leishmania* (7)



4.1.29

30.- *G. lamblia* (5), *D.*



4.1.30

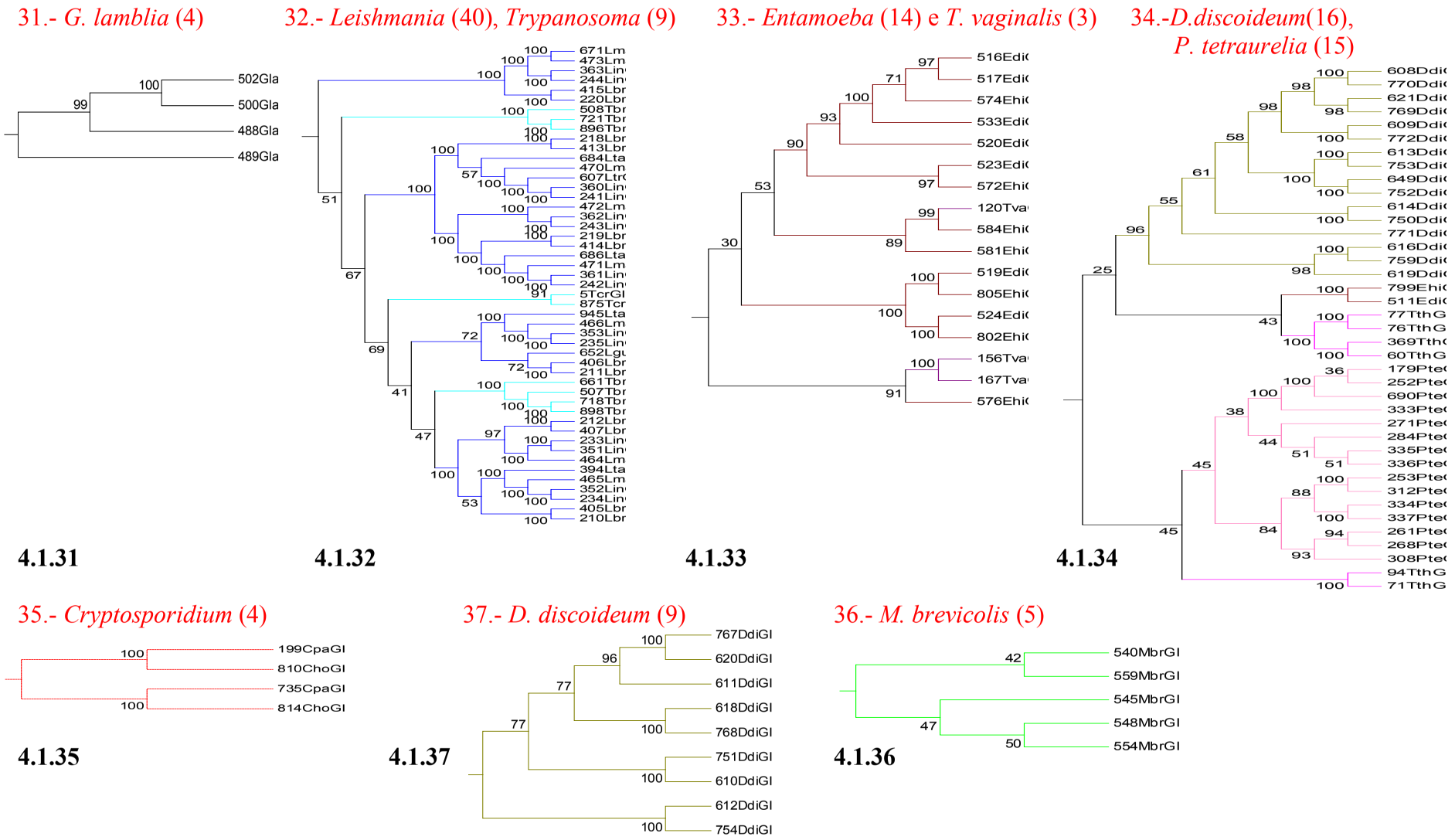


Figura 4.1 - Árvores filogenéticas AV construídas com PAUP para o gene MRPA

As árvores filogenéticas construídas segundo a Metodologia M1 para os genes candidatos de resistência a drogas. As árvores 4.1 representam ao gene MRPA.

4.1.2.2 Aquaporina - AQP

A árvore do gene AQP (104 espécies) apresentou os seguintes grupos monofiléticos mais representativos formados pelos gêneros: ***Plasmodium***, ***Paramecium***, ***Leishmania*** e ***Trypanosoma***. As espécies ***D. discoideum*** e ***M. brevicollis*** formaram o grupo basal.

Os gêneros *Leishmania* e *Trypanosoma* formaram dois grupos com as seguintes espécies: *L. major*, *L. infantum*, *L. brasiliensis* e *T. cruzi* (ou *T. brucei*). Por outro lado, *Leishmania* e *Trypanosoma*, apresentaram-se intimamente relacionados ao gênero *Plasmodium* e a duas sequências da espécie *T. gondii*. Outra terceira *T. gondii* posicionou-se relativamente mais perto ao grupo monofilético altamente distribuído do gênero *Paramecium*.

O gênero *Entamoeba* formou um grupo monofilético. *T. gondii* não formou um grupo bem resolvido (tabela 4.1, anexo 8.4).

4.1.2.3 Glicoproteína de Superfície de 63kDa - GP63

A árvore do gene GP63 (480 espécies) foi representada por grupos monofiléticos que apresentaram espécies ampla e diversificadamente distribuídas, apresentadas na tabela 4.1 (tabela 4.1, anexo 8.5).

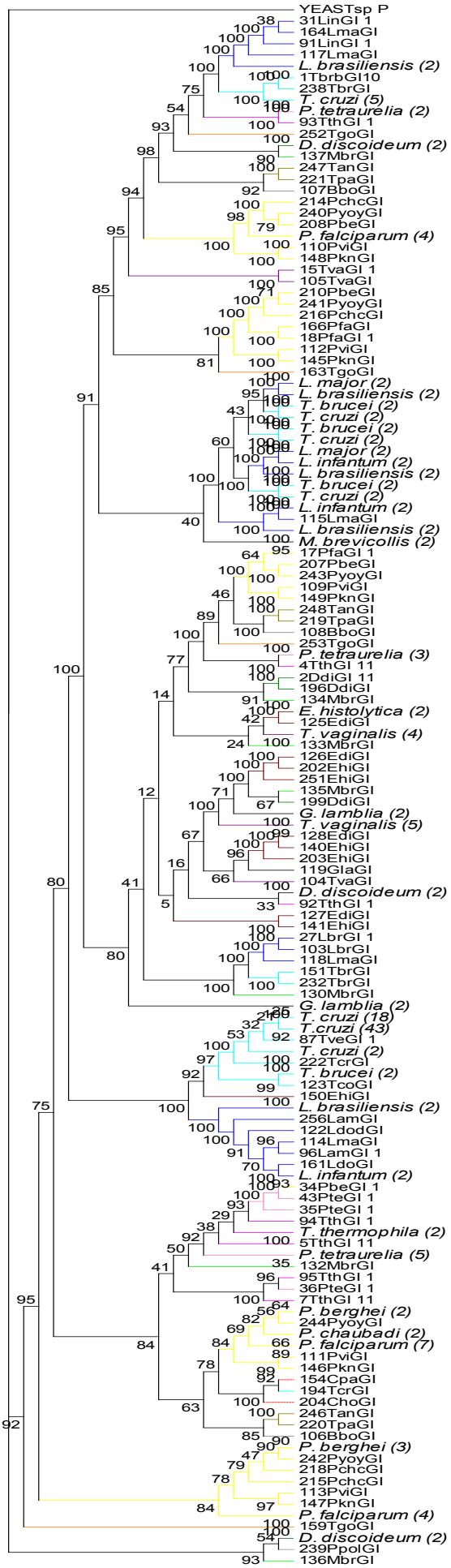
4.1.2.4 Proteínas de Choque Térmico de 70kDa - hsp70

A árvore do gene hsp70 (855 espécies) foi dividida em 14 grupos que se apresentaram como os mais resolvidos (tabela 4.3, anexo 8.6).

4.1.2.5 Tripanotiona Redutase - TRYR

A árvore do gene TRYR (256 espécies) foi representada pelas politomias compostas dos pequenos grupos de ***Trypanosoma***, ***Leishmania***, ***Entamoeba***, ***T. vaginalis*** e ***D. discoideum***. Todos os grupos foram muito diversificados. Esta árvore foi caracterizada pela proximidade filogenética entre *Trypanosoma* e *Leishmania*.

O gênero ***Plasmodium*** formou cinco subgrupos independentes formados pelas espécies *P. falciparum*, *P. berghei*, *P. vivax*, *P. knowlesi*, *P. chaubadi* e *P. yoelii*. Outros gêneros encontrados foram ***Theileria***, ***Cryptosporidium***, ***Tetrahymena***, ***Paramecium***, ***Giardia***, ***Toxoplasma***, ***Monosiga*** e ***Babesia*** (tabela 4.1, figura 4.2).



TRYR - TOTAL
PAUP - AV

Figura 4.2 - Árvores filogenéticas AV construídas com PAUP para o gene TRYR

As árvores filogenéticas construídas segundo a Metodologia M1 para os genes candidatos de resistência a drogas. A árvore 4.2 representa ao gene TRYR.

4.2

Tabela 4.3 - 14 grupos da árvore filogenética AV do gene hsp70. Metodologia M1

N. do Grupo	Nome do Grupo: Gênero / Espécie	N. de espécies	Espécies	Inferencias do grupo
2	*Leishmania* *Trypanosoma*	17 19	*L. brasiliensis*, *L. infantum*, *L. major*, *L. amazonensis*, *L. tarentolae*, *T. rangeli*, *T. cruzi* e *T. brucei*	Evento de duplicação formou dois subgrupos
5	**Vários** *Leishmania* *Trypanosoma*	87 8 26	*L. brasiliensis*, *L. infantum*, *L. major*, *T. cruzi* e *T. brucei*	Formou dois subgrupos separados por gênero. Apresenta outros três subgrupos só de espécies de *Leishmania*.
10	*Leishmania* *Trypanosoma*	13 25	*L. brasiliensis*, *L. infantum*, *L. major*, *L. amazonensis*, *L. chaubadi*, *L. tarentolae*, *T. cruzi*, *T. brucei* e *T. congolense*	Formaram subgrupos monofiléticos dos gêneros
11	*Leishmania* *Trypanosoma*	7 7	*L. brasiliensis*, *L. infantum*, *L. major*, *L. amazonensis*, *L. chaubadi*, *L. tarentolae*, *T. cruzi*, *T. brucei* e *T. congolense*	Formaram subgrupos monofiléticos dos gêneros
12	*Leishmania* *Trypanosoma*	4 6	*L. brasiliensis*, *L. infantum*, *L. major*, *L. amazonensis*, *L. chaubadi*, *L. tarentolae*, *T. cruzi*, *T. brucei* e *T. congolense*	Formaram subgrupos monofiléticos dos gêneros
13	*Leishmania* *Trypanosoma*	9 6	*L. brasiliensis*, *L. infantum*, *L. major*, *L. amazonensis*, *L. chaubadi*, *L. tarentolae*, *T. cruzi*, *T. brucei* e *T. congolense*	Formaram subgrupos monofiléticos dos gêneros
3	*Babesia*	40	Mais representativas: *B. canis*, *B. divergens* Outras: *B. bovis*, *B. caballi*, *B. odocoilei, B. sp* e *B. gibsoni*	Formaram dois subgrupos monofiléticos das espécies mais representativas
4	*Cryptosporidium*	141	Mais representativas: *C. parvum, C. meleagridis*, *C. hominis*, *C. baileyi*, *C. andersoni* e *C. felis* Outras: *C. sp, C. wrairi, C. serpentis, C. muris, C. andersoni e C. galli*	Formaram subgrupos monofiléticos das espécies mais representativas
6	*Entamoeba* *Trichomonas*	44 21	*E. dispar* e *E. histolytica*	Vários subgrupos em *Entamoeba*. *Trichomonas* formou intermediários
7	*Paramecium* *Tetrahymena*	18 7	*P. primaurelia, P. octaurelia, P. tetraurelia, P. sexaurelia, P. pentaurelia*, *P. novaurelia, P. tredecaurelia, P. jenningsi e T. thermophila*	Formaram subgrupos monofiléticos independentes dos gêneros mencionados
14	**Vários** *Paramecium*	 16	Mais representativa: *P. tetraurelia* Outros: *Babesia, Cryptosporidium, Tetrahymena e Theileria*	A espécie mais representativa formou monofilia
8	*Plasmodium*	12	Mais representativa: *P. falciparum* Outras: *P. berghei*, *P. yoelii*, *P. chabaudi*, *P. vivax* e *P. knowlesi*	A espécie mais representativa formou monofilia
9	*Plasmodium*	14	Mais representativa: *P. falciparum* Outras: *P. berghei*, *P. yoelii*, *P. chabaudi*, *P. vivax* e *P. knowlesi*	A espécie mais representativa formou monofilia
1	**Vários**	68	Várias espécies diferentes	Não formaram subgrupos determinados

4.1.3 Metodologia 2 (M2)

Ao comparar todas as árvores individuais dos melhores *hits* dos grupos iA, iB, iC, iiA, iiB e iiC - construídas com os diferentes algoritmos filogenéticos - em termos de valores de *bootstrap* ou probabilidades *aposteriori* para o MRBAYES - e consistência das relações filogenéticas e taxonômicas, as árvores mais consistentes foram obtidas com os algoritmos de MV (RAXML e PHYML) e IB (MRBAYES).

Os valores de *bootstrap* das árvores de MV e os números de geração das árvores de IB são mostrados, respectivamente, entre parênteses seguindo a ordem (RAXML-PHYML-MRBAYES). Optou-se por mostrar os nomes das espécies dos protozoários entre colchetes, para facilitar a relação visual com as topologias das árvores nas seções **Topologia** - contidas nas figuras - e **Topologia com comprimento de ramo -** contidas nos anexos.

As árvores construídas utilizando sequências completas e trimadas com o TRIMAL representaram melhor as relações filogenéticas e taxonômicas do que as árvores construídas com o GBLOCKS. Para o gene hsp70 as árvores não puderam ser construídas com o MRBAYES, devido a capacidade de memória insuficiente dos servidores, pelo alto alto custo computacional exigido e ao limite de execução do programa com um grande número de sequências.

A tabela 4.4 mostra os tamanhos dos alinhamentos para os grupos (i) todos os genes e (ii) quatro genes. O TRIMAL apresentou ser mais conservador que o GBLOCKS ao extrair um maior número de blocos conservados dos alinhamentos, com uma média de 22% para o TRIMAL e 7% para o GBLOCKS.

Tabela 4.4 - Tamanho dos alinhamentos completos e trimados com o TRIMAL e o GBLOCKS

Gene	Todos os genes (seis genes)				Quatro genes			
	N. sequências	Completo	TRIMAL	GBLOCKS	N. sequências	Completo	TRIMAL	GBLOCKS
MRPA	40	4.072	1.043 **(25%)**	0 **(-)**	30	4.072	1.018 **(25%)**	24 **(1%)**
AQP	23	2.607	303 **(12%)**	14 **(1%)**	22	2.607	303 **(12%)**	14 **(1%)**
GP63	23	1.694	505 **(30%)**	128 **(8%)**	13	1.694	538 **(32%)**	131 **(8%)**
hsp70	103	1.542	124 **(8%)**	0 (-)	29	1.542	120 **(8%)**	0 **(-)**
TRYR	35	816	446 **(55%)**	81 **(10%)**	28	816	445 **(55%)**	24 **(3%)**

4.1.3.1 MRPA

(i) Árvores construídas com as espécies que apresentaram o melhor *hit* em todos os genes

Na tabela 4.5, as árvores do gene **MRPA do grupo iA** - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos - apresentaram três grupos mais representativos (*Leishmania*, *Plasmodium*, *Trypanosoma-Leishmania*) (figura 4.3.1, anexo 8.7.1).

O PHYML mostrou *Leishmania*, *Trypanosoma* e *Plasmodium* com os comprimentos de ramo mais curtos nos grupos intra-gênero. *Entamoeba*, *Tetrahymena*, *Trichomonas*, *Toxoplasma*, *Monosiga*, *Giardia* e *Paramecium* tiveram os maiores comprimentos de ramos (anexo 8.7.24, 8.7.25).

As árvores do gene **MRPA do grupo iB** - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o GBLOCKS 0.91b - não puderam ser construídas com o GBLOCKS, uma vez que o programa não retornou blocos conservados.

(ii) Árvores construídas com as espécies que apresentaram o melhor *hit* em quatro genes

As topologias das árvores do gene **MRPA dos grupos:**

- iiA - espécies apresentando o melhor *hit* em quatro genes / sequências de genes completos - (figura 4.3.3, anexo 8.7.9-8.7.13),

- iiB - espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o GBLOCKS 0.91b - (figura 4.3.4, anexo 8.7.14-8.7.18) e

- iiC - espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o TRIMAL v1.2 - (figura 4.3.5, anexo 8.7.19-8.7.23) foram muito semelhantes às árvores do gene **MRPA dos grupos:**

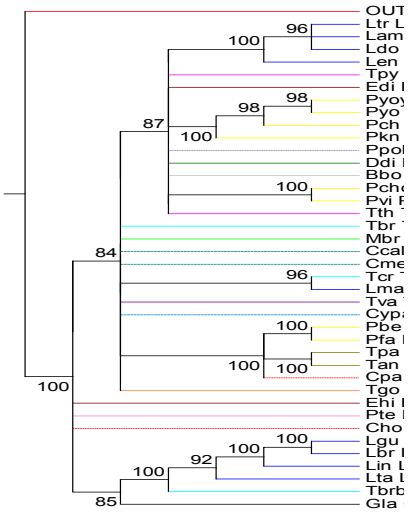
- iA - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos (figura 4.3.1, anexo 8.7.1-8.7.4),

- iC - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o TRIMAL v1.2 , respectivamente (figura 4.3.2, anexo 8.7.5-8.7.8).

Tabela 4.5 - Árvores construídas com as espécies que apresentaram o melhor *hit* em todos os genes. Metodologia M2

Gene	Nome do Grupo: Gênero	Valor de consistência	Espécies	Observações do grupo
MRPA	*Leishmania*	100-100	*L. tropica*, *L. donovani*, *L. amazonensis*, *L. enriettii*	
	Plasmodium	100-100	*P. yoelii yoelii*, *P. yoelii*, *P. knwolesi*, *P. chaubadi*	Árvores **MRPA - Todas as Sequências Completas**
	Trypanosoma-Leishmania	85-98	*L. braziliensis, L. guyanensis, L. infantum*, *L. tarentolae*, *T. brucei brucei*	
AQP	*Leishmania*	94-100-99	*L. donovani*, *L. infantum*, *L. mexicana*, *L. major*, *L. tarentolae*	
	Trypanosoma	99-100-100	*T. brucei, T. brucei brucei*	Árvores **AQP - Todas as Sequências Completas**
	Plasmodium	100-100-99	*P. berghei*, *P. yoelii yoelii*, *P. knwolesi*, *P. vivax*, *P. falciparum*	
	Entamoeba	100-100-100	*E. histolytica*, *E. díspar*	
GP63	*Leishmania*	100-100-100	*L. aethiopica*, *L. tropica*, *L. turanica*, *L. arabica*, *L. major*, *L. donovani, L. sp*, *L. mexicana*, *L. donovani*, *L. infantum*, *L. amazonensis*, *L. braziliensis, L. guyanensis, L. panamensis*	Árvores **GP63 - Todas as Sequências Completas**
	Trypanosoma	91-100-100	*T. cruzi, T. brucei, T. brucei rangeli*	
	Entamoeba	100-100-100	*E. histolytica*, *E. díspar*	
hsp70	*Cryptosporidium*	100-*	Vários	
	Babesia, Theileria	100-85	Vários	Árvores **hsp70 - Todas as Sequências Completas**
	Leishmania, *Trypanosoma*	100-99	Vários	
	Paramecium	94-*	Vários	
TRYR	*Leishmania*	90-100-99	Vários	
	Trypanosoma	98-99-97	Vários	Árvores **TRYR - Todas as Sequências Completas**
	Plasmodium	100-100-99	*P. berghei*, *P. yoelii yoelii*, *P. knwolesi*, *P. vivax*, *P. falciparum, P. chaubadi chaubadi*	

MRPA - Todas as Sequências Completas
Topologia - RAXML



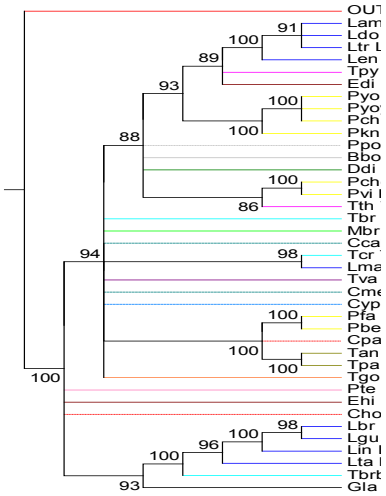
4.3.1

Trimadas com o GBLOCKS

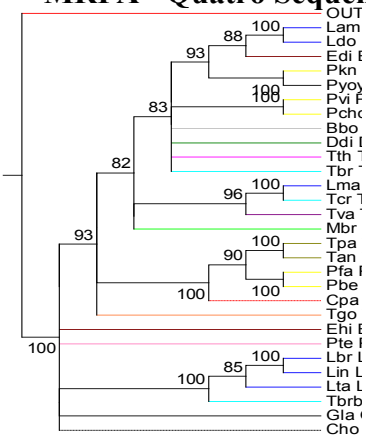
Árvore não obtida - o GBLOCKS não retornou bloco conservado -

4.3.2

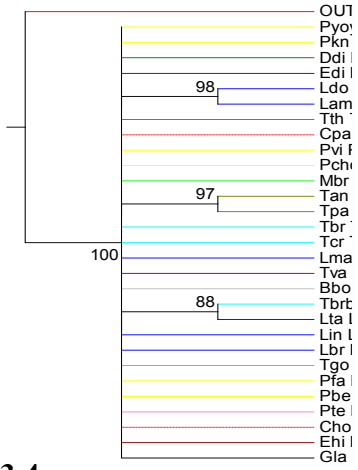
Trimadas com o TRIMAL



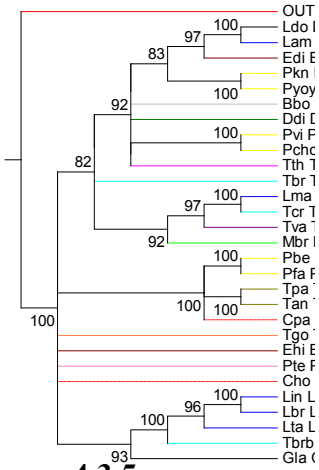
MRPA - Quatro Sequências



4.3.3



4.3.4



4.3.5

Figura 4.3 - MRPA- Árvores filogenéticas construídas com os programas RAXML
As árvores filogenéticas do gene MRPA foram construídas segundo a Metodologia M2 para os genes candidatos de resistência a drogas.

Todas as Sequências Completas:
- a *árvore* 4.3.1 representa o RAXML

Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 4.3.2 representa o RAXML

Quatro Sequências Completas:
- a *árvore* 4.3.3 representa o RAXML

Quatro Sequências Trimadas com o GBLOCKS:
- a *árvore* 4.3.4 representa o RAXML,

Quatro Sequências Trimadas com o TRIMAL:
- *a árvore* 4.3.5 representa o RAXML

4.1.3.2 Aquaporina - AQP

(i) Árvores construídas com as espécies que apresentaram o melhor *hit* em todos os genes

Na tabela 4.5, as árvores do gene **AQP do grupo iA** - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos - formaram grupos monofiléticos com valores de *bootstrap* consistentes, para todos os algoritmos filogenéticos (figura 4.4.1, anexo 8.8.1-8.8.5).

As árvores filogenéticas construídas com o PHYML (anexo 8.8.1) e o MRBAYES (anexo 8.8.2) foram muito similares. As árvores construídas com o PHYML (anexo 8.8.31) tenderam a formar politomia. *Trypanosoma*, *Leishmania*, *Plasmodium* e *Entamoeba* tiveram comprimentos de ramo curtos. Comprimentos de ramos mais compridos foram apresentados por *T. thermophila*, *P. tetraurelia*, *M. brevicollis* e *T. gondii* (anexo 8.8.31-8.8.42).

As topologias das árvores do gene **AQP dos grupos:**

- iA - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos - (figura 4.4.1, anexo 8.8.1-8.8.5) e

- iC - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o TRIMAL v1.2 - (figura 4.4.3, anexo 8.8.11-8.8.15) foram idênticas e estas foram muito semelhantes às arvores **do grupo:**

- iB - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o GBLOCKS 0.91b - (figura 4.4.2, anexo 8.8.6-8.8.10). Todas as topologias formaram os mesmos grupos monofiléticos.

Todas as árvores construídas com os diferentes programas apresentaram topologias similares. As árvores PAUP-MP e PAUP-AV mantiveram os mesmos grupos monofiléticos. Uma exceção foi encontrada nas árvores PHYML (anexo 8.8.1, 8.8.6, 8.8.11) que tenderam a formar politomias e a outra exceção foi na árvore MRBAYES usando sequências trimadas com o GBLOCKS (anexo 8.8.7) que foi a mais diferente e apresentou tendência a formar politomias.

(ii) Árvores construídas com as espécies que apresentaram o melhor *hit* em quatro genes

As árvores do gene **AQP dos grupos:**

- iiA - espécies apresentando o melhor *hit* em quatro genes / sequências de genes completos - (figura 4.4.4, anexo 8.8.16-8.8.20),

- iiB - espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o GBLOCKS 0.91b - (figura 4.4.5, anexo 8.8.21-8.8.25) e

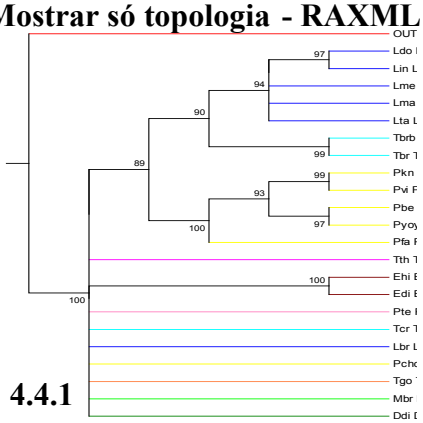
- iiC - espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o TRIMAL v1.2 - (figura 4.4.6, anexo 8.8.26-8.8.30) mantiveram características muito semelhantes em termos de topologia e valores de *bootstrap* com as árvores **dos grupos:**

- iA (figura 4.4.1, anexo 8.8.1-8.8.5),

- iB (figura 4.4.2, anexo 8.8.6-8.8.10) e

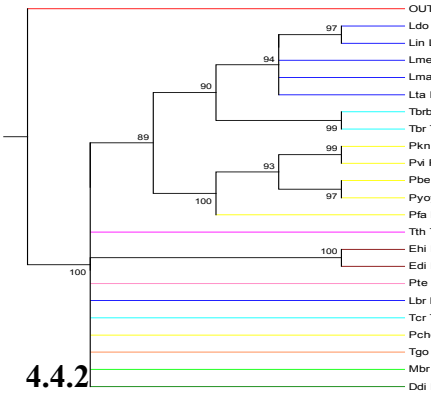
- iC (figura 4.4.3, anexo 8.8.11-8.8.15), respectivamente.

AQP - Todas as Sequências Completas
Mostrar só topologia - RAXML



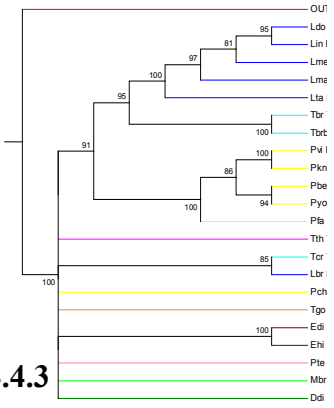
4.4.1

GBLOCKS



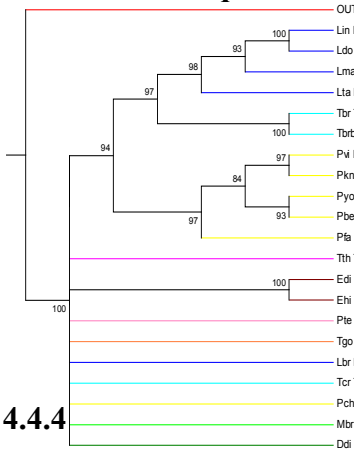
4.4.2

TRIMAL

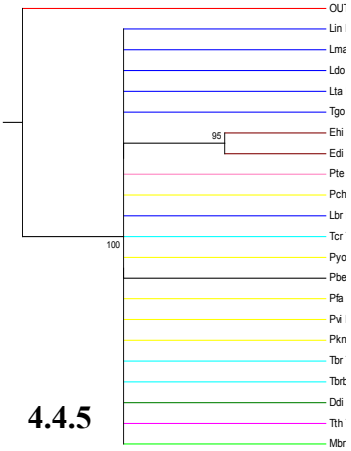


4.4.3

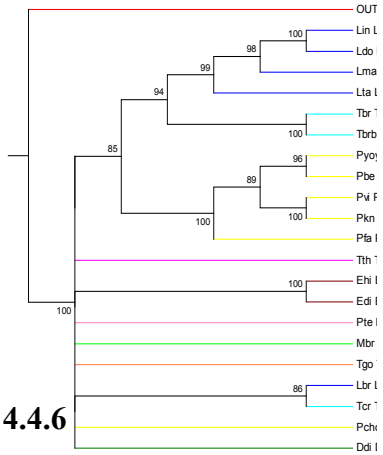
AQP - Quatro Sequências



4.4.4



4.4.5



4.4.6

Figura 4.4 - AQP - Árvores filogenéticas construídas com os programas RAXML

As árvores filogenéticas do gene AQP foram construídas segundo a Metodologia M2 para os genes candidatos de resistência a drogas.

Todas as Sequências Completas:
- a *árvore* 4.4.1 representa o RAXML

Todas as Sequências Trimadas com o GBLOCKS:
- a *árvore* 4.4.2 representa o RAXML

Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 4.4.3 representa o RAXML

Quatro Sequências Completas:
- a *árvore* 4.4.4 representa o RAXML

Quatro Sequências Trimadas com o GBLOCKS:
- a *árvore* 4.4.5 representa o RAXML

Quatro Sequências Trimadas com o TRIMAL:
- a *árvore* 4.4.6 representa o RAXML

4.1.3.3 Glicoproteína de Superfície de 63kDa - GP63

(i) Árvores construídas com as espécies que apresentaram o melhor *hit* em todos os genes

Na tabela 4.5, as árvores do gene **GP63 do grupo iA** - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos - formaram grupos monofiléticos com valores de *bootstrap* consistentes, para todos os algoritmos filogenéticos (figura 4.5.1, anexo 8.9.1-8.9.5).

As árvores filogenéticas construídas com o PHYML (anexo 8.9.1) e o MRBAYES (anexo 8.9.2) foram muito similares. O PHYML tendeu a formar politomias.

Leishmania teve os comprimentos de ramo mais curtos nos grupos intra-gênero em comparação a *Trypanosoma* e a outros protozoários (anexo 8.9.31-8.9.42).

As topologias das árvores do gene **GP63 dos grupos:**

- iA - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos - (figura 4.5.1, anexo 8.9.1-8.9.5) e

- iC - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o TRIMAL v1.2 - (figura 4.5.2, anexo 8.9.11-8.9.15) foram idênticas. Já quando comparadas às **do grupo:**

- iB - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o GBLOCKS 0.91b - (figura 4.5.1, anexo 8.9.6-8.9.10) foram muito semelhantes. As árvores do grupo **iB**, tenderam a formar politomias com os programas PHYML (anexo 8.9.6) e MRBAYES (anexo 8.9.7).

(ii) Árvores construídas com as espécies que apresentaram o melhor *hit* em quatro genes

As árvores do gene **GP63 dos grupos:**

- iiA - espécies apresentando o melhor *hit* em quatro genes / sequências de genes completos - (figura 4.5.4, anexo 8.9.16-8.9.20),

- iiB - espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o GBLOCKS 0.91b (figura 4.5.5, anexo 8.9.21-8.9.25) e

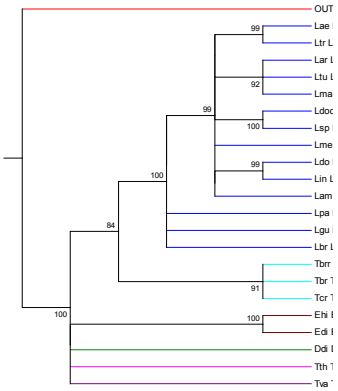
- iiC - espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o TRIMAL v1.2 (figura 4.5.6, anexo 8.9.26-8.9.30) mantiveram características muito semelhantes em termos de topologia e valores de *bootstrap* com as árvores do gene **GP63 dos grupos:**

- iA (figura 4.5.1, anexo 8.9.1-8.9.5),

- **iB** (figura 4.5.1, anexo 8.9.6-8.9.10) e

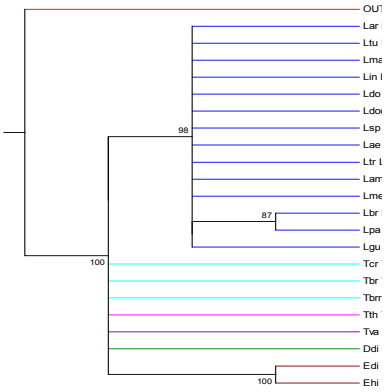
- **iC** (figura 4.5.2, anexo 8.9.11-8.9.15), respectivamente.

GP63 - Todas as Sequências Completas
Mostrar só topologia - RAXML



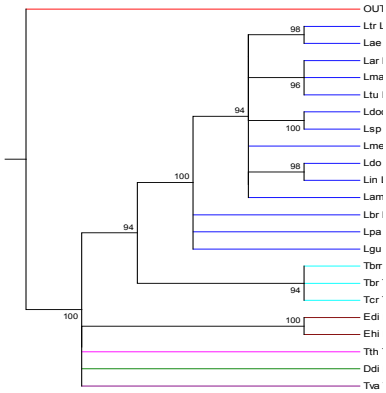
4.5.1

GBLOCKS



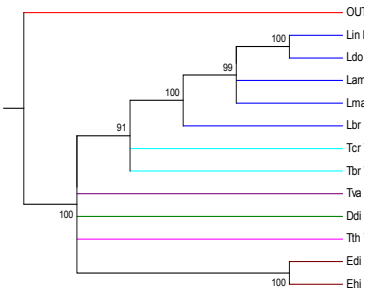
4.5.2

TRIMAL

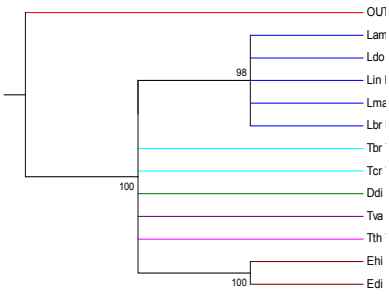


4.5.3

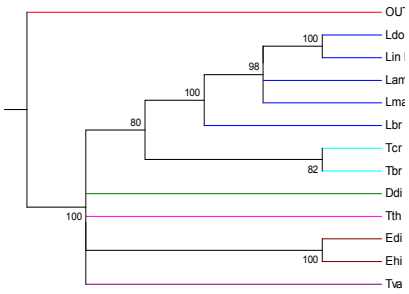
GP63 - Quatro Sequências



4.5.4



4.5.5



4.5.6

Figura 4.5 - GP63 - Árvores filogenéticas construídas com os programas RAXML

As árvores filogenéticas do gene GP63 foram construídas segundo a Metodologia M2 para os genes candidatos de resistência a drogas.

Todas as Sequências Completas:
- a *árvore* 4.5.1 representa o RAXML

Todas as Sequências Trimadas com o GBLOCKS:
- a *árvore* 4.5.2 representa o RAXML

Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 4.5.3 representa o RAXML

Quatro Sequências Completas:
- a *árvore* 4.5.4 representa o RAXML

Quatro Sequências Trimadas com o GBLOCKS:
- a *árvore* 4.5.5 representa o RAXML

Quatro Sequências Trimadas com o TRIMAL:
- a *árvore* 4.5.6 representa o RAXML

4.1.3.4 Proteínas de Choque Térmico de 70kDa - hsp70

(i) Árvores construídas com as espécies que apresentaram o melhor hit em todos os genes

Na tabela 4.5, as árvores do gene **hsp70 do grupo iA** - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos - formaram grupos monofiléticos com valores de *bootstrap* consistentes, para todos os algoritmos filogenéticos (figura 4.6.1, anexo 8.10.1-8.10.4).

A topologia das árvores do gene **hsp70 dos grupos:**

- iA - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos - (figura 4.6.1, anexo 8.10.1-8.10.4) e

- iC - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o TRIMAL v1.2 - (figura 4.6.2, anexo 8.10.5-8.10.8) foram idênticas.

As árvores filogenéticas construídas com o PHYML (anexo 8.10.1, 8.10.5) tenderam a formar politomias. As distâncias filogenéticas medidas com o PHYML (anexo 8.10.17, 8.10.18) mostraram que todas as espécies tiveram os comprimentos de braços curtos. Comprimentos de ramos mais compridos foram obtidos nas espécies *E. dispar*, *B. natans*, *G. lamblia* e *P. berghei*.

As árvores do gene **hsp70 do grupo iB** - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o GBLOCKS 0.91b -, não puderam ser construídas. O GBLOCKS não retornou nenhum bloco conservado para o hsp70.

(ii) Árvores construídas com as espécies que apresentaram o melhor *hit* em quatro genes

As árvores do gene **hsp70 dos grupos:**

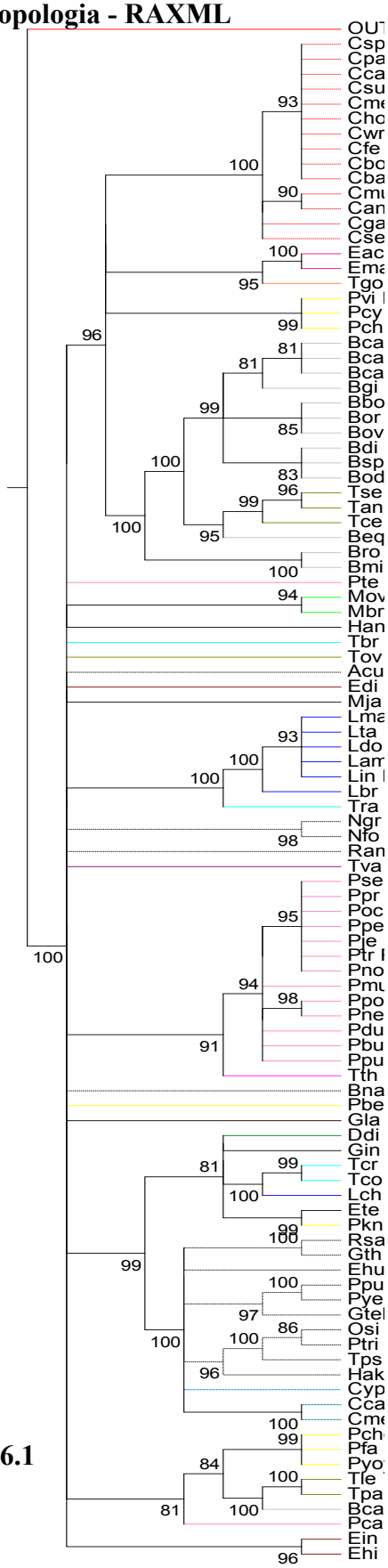
- iiA - espécies apresentando o melhor *hit* em quatro genes / sequências de genes completos - (figura 4.6.3, anexo 8.10.9-8.10.12) e

- iiC - espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o TRIMAL v1.2 - (figura 4.6.4, anexo 8.10.13-8.10.16) mostraram similares topologias e valores de *bootstrap*. Os grupos mais representativos foram ***Leishmania*** e ***Plasmodium***. Estas árvores mostraram diferentes topologias quando comparadas com as árvores do gene **hsp70 dos grupos:**

- iA (figura 4.6.1, anexo 8.10.1-8.10.4) e

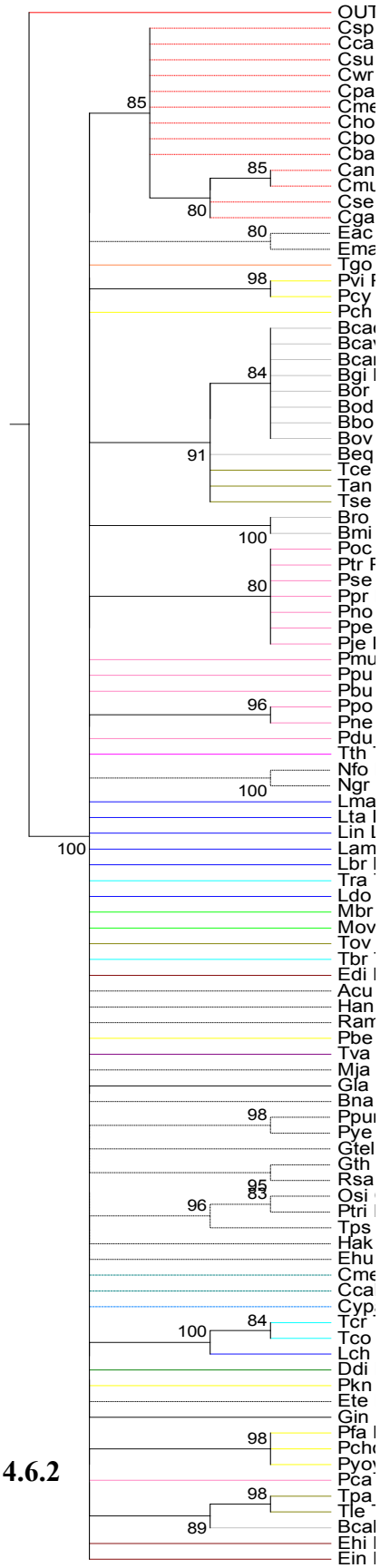
- iC (figura 4.6.2, anexo 8.10.5-8.10.8), respectivamente.

hsp70 - Todas as Sequências Completas
Topologia - RAXML



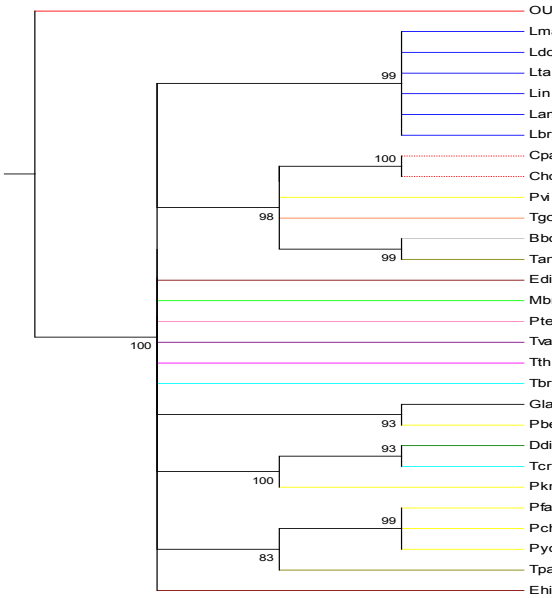
4.6.1

TRIMAL



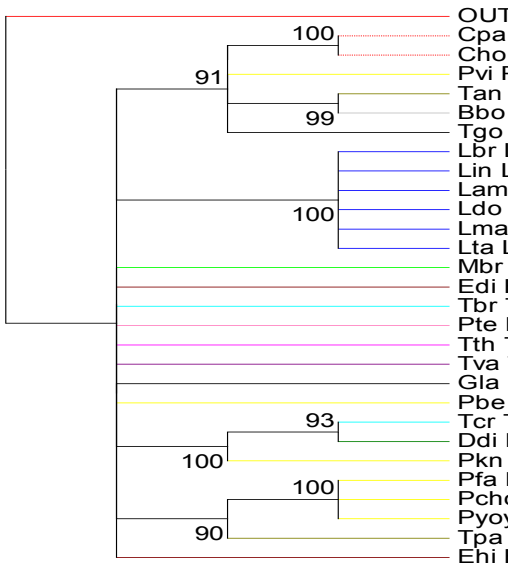
4.6.2

hsp70 - Quatro Sequências Completas
Mostrar só topologia - RAXML



4.6.3

TRIMAL



4.6.4

Figura 4.6 - hsp70 - Árvores filogenéticas construídas com os programas RAXML

As árvores filogenéticas do gene hsp70 foram construídas segundo a Metodologia M2 para os genes candidatos de resistência a drogas.

Todas as Sequências Completas:
- a *árvore* 4.6.1 representa o RAXML

Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 4.6.2 representa o RAXML

Quatro Sequências Completas:
- a *árvore* 4.6.3 representa o RAXML

Quatro Sequências Trimadas com o TRIMAL:
- a *árvore* 4.6.4 representa o RAXML

4.1.3.5 Tripanotiona Redutase - TRYR

(i) Árvores construídas com as espécies que apresentaram o melhor *hit* em todos os genes

Na tabela 4.5, as árvores do gene **TRYR do grupo iA** - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos - formaram grupos monofiléticos com valores de *bootstrap* consistentes, para todos os algoritmos filogenéticos (figura 4.7.1, anexo 8.11.1-8.11.5).

As árvores filogenéticas construídas com o PHYML (anexo 8.11.1) e o MRBAYES (anexo 8.11.2) foram muito similares. *Leishmania*, *Trypanosoma*, *Plasmodium* tiveram os comprimentos de ramo mais curtos nos grupos intra-gênero. *Giardia* e *Cryptosporidium* apresentaram comprimentos de ramo muito semelhantes. *Entamoeba*, *Giardia*, *Monosiga*, *Trichomonas*, *Toxoplasma*, *Tetrahymena* e *Paramecium* tiveram comprimentos de ramos mais compridos (anexo 8.11.31-8.11.42).

As topologias das árvores do gene **TRYR dos grupos:**

- iA - espécies apresentando o melhor *hit* em todos os genes / sequências de genes completos - (figura 4.7.1, anexo 8.11.1-8.11.5) e

- iC - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o TRIMAL v1.2 - (figura 4.7.3, anexo 8.11.11-8.11.15) foram idênticas e muito semelhantes às árvores **do grupo:**

- iB - espécies apresentando o melhor *hit* em todos os genes / sequências de genes trimados com o GBLOCKS 0.91b - (figura 4.7.2, anexo 8.11.6-8.11.10), exceto porque estas últimas apresentaram valores de *bootstrap* menores.

(ii) Árvores construídas com as espécies que apresentaram o melhor *hit* em quatro genes

As árvores do gene **TRYR dos grupos:**

- iiA - espécies apresentando o melhor *hit* em quatro genes / sequências de genes completos - (figura 4.7.4, anexo 8.11.16-8.11.20),

- iiB - espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o GBLOCKS 0.91b - (figura 4.7.5, anexo 8.11.21-8.11.25) e

- iiC - espécies apresentando o melhor *hit* em quatro genes / sequências de genes trimados com o TRIMAL v1.2 - (figura 4.7.6, anexo 8.11.26-8.11.30) foram muito semelhantes às árvores do gene **TRYR dos grupos:**

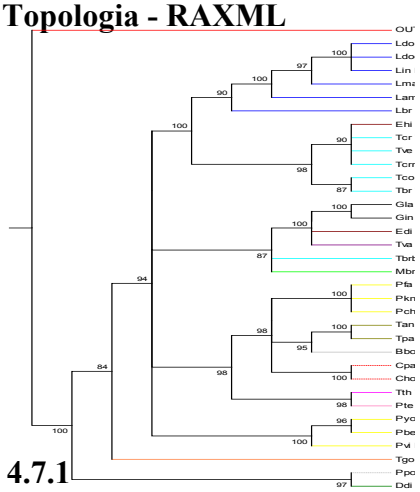
- **iA** (figura 4.7.1, anexo 8.11.1-8.11.5),

- **iB** (figura 4.7.2, anexo 8.11.6-8.11.10) e

- **iC** (figura 4.7.3, anexo 8.11.11-8.11.15), respectivamente.

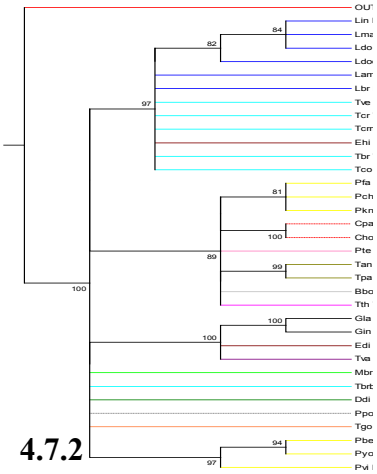
TRYR - Todas as Sequências Completas

Topologia - RAXML



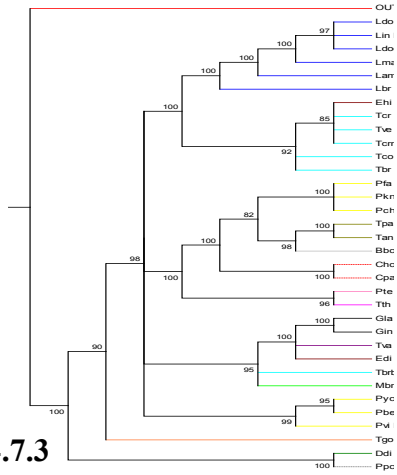
4.7.1

GBLOCKS



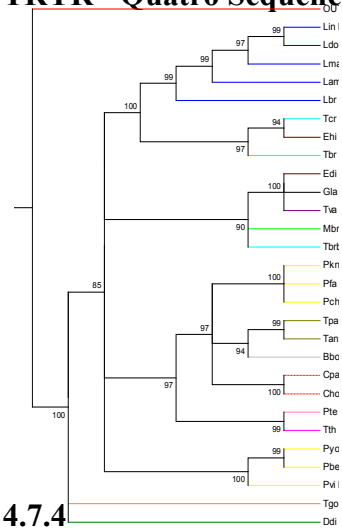
4.7.2

TRIMAL

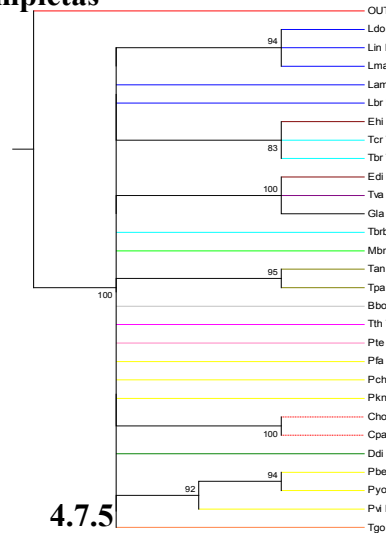


4.7.3

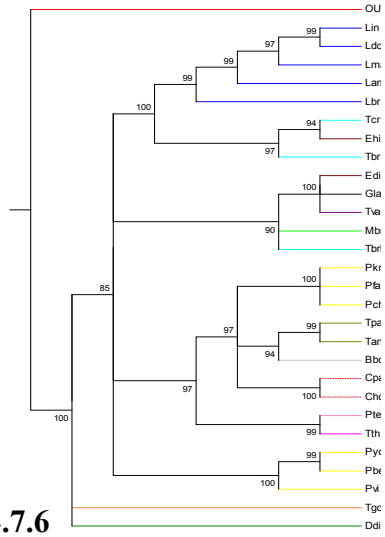
TRYR - Quatro Sequências Completas



4.7.4



4.7.5



4.7.6

Figura 4.7 - TRYR - Árvores filogenéticas construídas com os programas RAXML

As árvores filogenéticas do gene TRYR foram construídas segundo a Metodologia M2 para os genes candidatos de resistência a drogas.

Todas as Sequências Completas:
- a *árvore* 4.7.1 representa o RAXML

Todas as Sequências Trimadas com o GBLOCKS:
- a *árvore* 4.7.2 representa o RAXML

Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 4.7.3 representa o RAXML

Quatro Sequências Completas:
- a *árvore* 4.7.4 representa o RAXML

Quatro Sequências Trimadas com o GBLOCKS:
- a *árvore* 4.7.5 representa o RAXML

Quatro Sequências Trimadas com o TRIMAL:
- a *árvore* 4.7.6 representa o RAXML

4.1.4 Metodologia 3 (M3)

A árvore da supermatriz (anexo 8.12.1) dos Genes de Resistência a Drogas em Protozoários foi formada pelos seguintes grupos representativos: ***Paramecium***, ***Cryptosporidium***, ***Babesia***, ***Leishmania*** e ***Plasmodium***. Grupos similares, mas incompletos, foram obtidos com a superárvore (anexo 8.12.2).

4.2 Filogenômica baseada na reconstrução da árvore de espécies de protozoários

No final, foram obtidos 2 alinhamentos concatenados (1 total de 21.260 posições e 1 trimado de 12.807 posições) e 62 alinhamentos individuais (31 totais e 31 trimados) em 74 espécies de protozoários.

4.2.1 Teste do sinal filogenético

A presença do sinal filogenético nos ortólogos universais foi demonstrada com o Teste PTP e a Estatística G1. Os 64 alinhamentos passaram no Testes PTP e a Estatística G1, confirmando a presença do sinal filogenético em todos os alinhamentos.

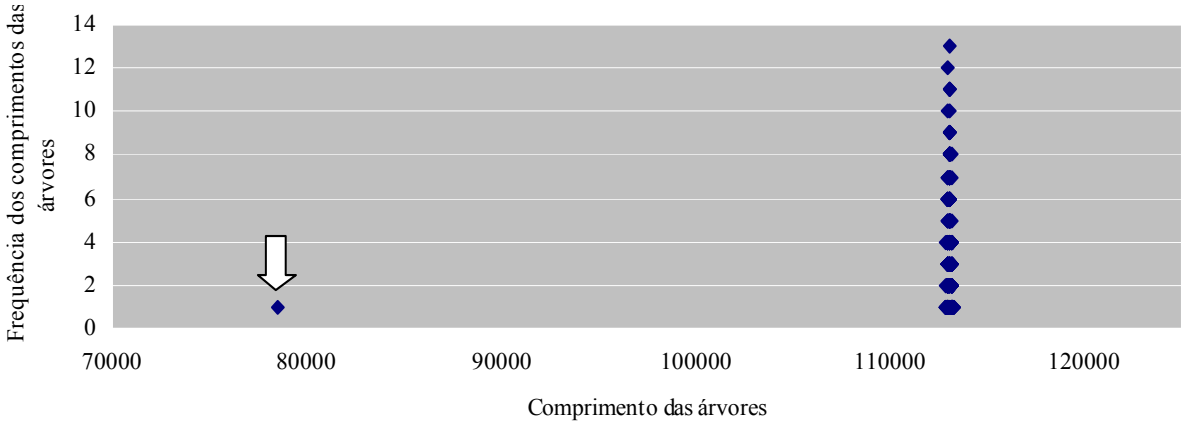
O Teste PTP, executado sob os alinhamentos concatenados **M1 (total)** (85268 passos, $p < 0.001$) (Figura 4.8B) e **M2 (trimados)** (78498 passos, $p < 0.001$) (Figura 4.8A), indicou que o valor do número de passos, relacionado à distância da árvore mais parcimoniosa obtida com os dados originais, está longe em comparação aos valores das outras árvores obtidas através da permutação dos dados.

A Estatística G1 mostrou que a distribuição das árvores foi devidamente inclinada (em relação a uma distribuição normal) na direção negativa para **M1 (total)** ($G1=-0,58$, $p < 0.001$) (Figura 4.9B) e para **M2 (trimados)** ($G1=-0,57$, $p < 0.001$) (figura 4.9A), isto é de acordo com a assimetria da curva da árvore mais parcimoniosa ocorrendo à esquerda de todas as outras árvores randomizadas.

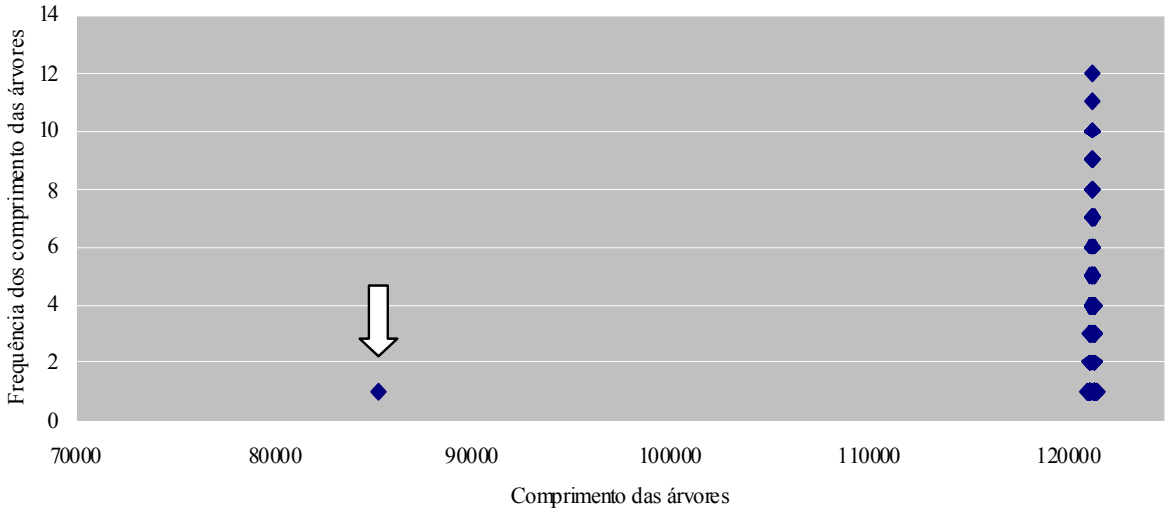
Outros resultados do Teste PTP com relação aos caracteres dos alinhamentos concatenados são:

(I) Caracteres Totais: M1 (total) 21.260 *versus* a média do alinhamento simples 686 e **M2 (trimados)** 12.807 *versus* a média do alinhamento simples 412.

(II) Caracteres Parcimoniosos Informativos: M1 (total) 9.308 *versus* a média do alinhamento simples 300 e **M2 (trimados)** 8.586 *versus* a média do alinhamento simples 277 (tabelas 4.6, 4.7).



A

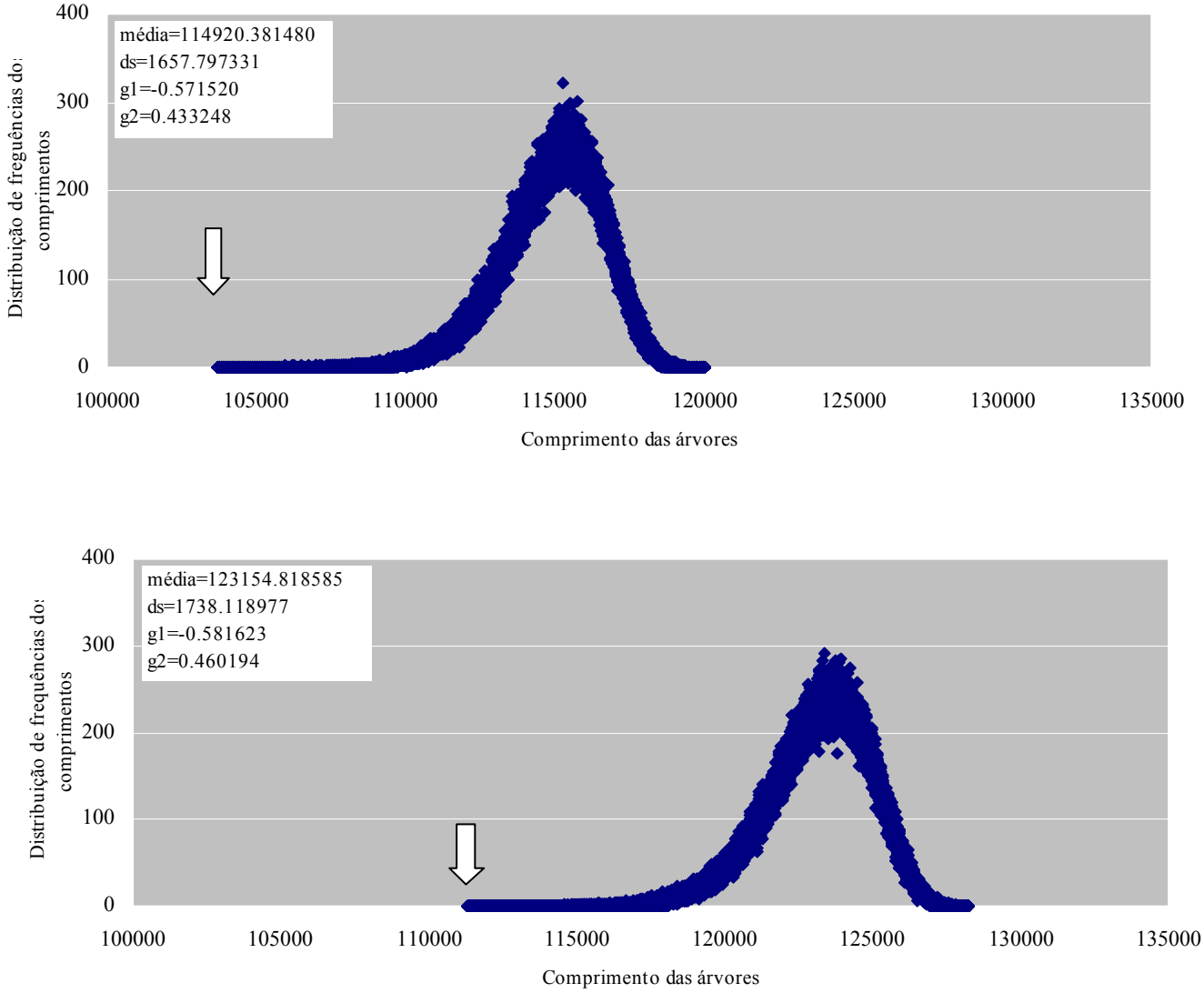


B

Figura 4.8- Gráfico de distribuição do Teste de Permutação (PTP) dos ortólogos universais em protozoários

A. Teste PTP dos alinhamentos M2 (trimados). (Número de réplicas=1000, busca=heurística). Seta cinza indica a árvore mais parcimoniosa (AMP=78498). ($p < 0.001$) indica diferença significativa entre os dados originais da AMP e os dados permutados.

B. Teste PTP dos alinhamentos M1 (total). (Número de réplicas=1000, busca=heurística). Seta cinza indica a árvore mais parcimoniosa (AMP=85268). ($p < 0.001$) indica diferença significativa entre os dados originais da AMP e os dados permutados.



B

Figura 4.9- Gráfico da Estatística G1 dos ortólogos universais em protozoários

A. Estatística G1 dos alinhamentos M2 (trimados). (Número de réplicas=1000000). Seta cinza indica a árvore mais parcimoniosa (AMP=103691). ($p < 0.001$) indica diferença significativa entre os dados originais da AMP e os dados permutados.
B. Estatística G1 dos alinhamentos M1 (total). (Número de réplicas=1000000). Seta Estatística indica a árvore mais parcimoniosa (AMP=111342). ($p < 0.001$) indica diferença significativa entre os dados originais da AMP e os dados permutados.

Tabela 4.6 - Resultados do Teste PTP dos dois alinhamentos concatenados dos ortólogos universais em protozoários
Resumo do estado dos caracteres, critério=parcimônia (PAUP).

Alinhamentos concatenados dos ortólogos universais	Caracteres Totais	Caracteres Constantes	Caracteres Parcimoniosos No Informativos Variáveis	Caracteres Parcimoniosos Informativos
Trimado	12.807	1.836	2.385	8.586
Total	21.260	7.479	4.473	9.308

Tabela 4.7 - Resultados do Teste PTP dos 31 alinhamentos simples dos ortólogos universais em protozoários.
Resumo do estado dos caracteres, critério=parcimônia (PAUP).

Alinhamentos dos ortólogos universais		Caracteres Totais	Caracteres Constantes	Caracteres Parcimoniosos No Informativos Variáveis	Caracteres Parcimoniosos Informativos
COG0012	Total	605	153	59	393
	Trimado	363	11	14	338
COG0016	Total	964	249	228	487
	Trimado	578	44	59	475
COG0048	Total	241	49	58	134
	Trimado	145	9	6	130
COG0049	Total	292	103	17	172
	Trimado	175	6	12	157
COG0052	Total	339	18	23	298
	Trimado	203	2	5	196
COG0080	Total	1543	1223	132	188
	Trimado	926	606	132	188
COG0081	Total	477	55	112	307
	Trimado	286	2	14	270
COG0087	Total	615	115	145	355
	Trimado	369	16	29	324
COG0091	Total	522	348	27	147
	Trimado	313	139	27	147
COG0092	Total	338	77	38	223
	Trimado	203	5	7	191
COG0093	Total	256	84	26	146
	Trimado	154	2	6	146
COG0094	Total	242	10	32	200
	Trimado	148	2	5	141
COG0096	Total	170	28	16	126

	Trimado	120	4	2	114
COG0097	Total	295	75	26	194
	Trimado	177	1	2	174
COG0098	Total	298	27	32	239
	Trimado	179	11	8	160
COG0099	Total	249	84	21	144
	Trimado	149	5	10	134
COG0100	Total	274	114	32	128
	Trimado	164	9	27	128
COG0102	Total	805	467	109	229
	Trimado	438	145	109	229
COG0103	Total	738	85	421	232
	Trimado	443	5	206	232
COG0172	Total	2.065	1.033	522	510
	Trimado	1.239	207	522	510
COG0184	Total	540	330	82	128
	Trimado	324	129	67	128
COG0186	Total	320	122	45	153
	Trimado	192	21	18	153
COG0197	Total	234	31	8	195
	Trimado	140	5	3	132
COG0200	Total	666	161	287	218
	Trimado	400	6	176	218
COG0201	Total	572	71	69	432
	Trimado	374	15	12	347
COG0202	Total	1.033	373	271	389
	Trimado	620	24	207	389
COG0256	Total	704	346	78	280
	Trimado	422	64	78	280
COG0495	Total	2307	521	629	1.157
	Trimado	1.384	27	226	1.131
COG0522	Total	880	542	79	259
	Trimado	528	190	79	259
COG0525	Total	1.611	311	396	904
	Trimado	967	67	76	824
COG0533	Total	1065	274	450	341
	Trimado	639	57	241	341

4.2.2 Análises filogenéticas

A tabela 4.8 mostra as matrizes dos modelos evolutivos para os alinhamentos totais (M1) , trimados (M2) e para os grupos G1, G2 e G3.

Tabela 4.8 - Matrizes dos modelos evolutivos

Ortólogos Universais	Matriz para os alinhamentos totais	Matriz para os alinhamentos trimados
COG0012	Blosum62	RtREV
COG0016	Blosum62	WAG
COG0048	RtREV	RtREV
COG0049	RtREV	RtREV
COG0052	Blosum62	RtREV
COG0080	RtREV	RtREV
COG0081	JTT	RtREV
COG0087	WAG	RtREV
COG0091	RtREV	RtREV
COG0092	CPREV	CPREV
COG0093	WAG	RtREV
COG0094	RtREV	RtREV
COG0096	RtREV	RtREV
COG0097	WAG	RtREV
COG0098	WAG	WAG
COG0099	RtREV	RtREV
COG0100	CPREV	RtREV
COG0102	CPREV	WAG
COG0103	JTT	RtREV
COG0172	WAG	RtREV
COG0184	RtREV	RtREV
COG0186	Blosum62	WAG
COG0197	RtREV	RtREV
COG0200	Blosum62	WAG
COG0201	Blosum62	WAG
COG0202	Blosum62	WAG
COG0256	WAG	WAG
COG0495	Blosum62	RtREV
COG0522	Blosum62	RtREV
COG0525	Blosum62	RtREV
COG0533	Blosum62	RtREV
Grupo G1	Blosum62	Blosum62
Grupo G2	RtREV	RtREV
Grupo G3	WAG	WAG

A figura 4.10 mostra a comparação das árvores usando as sequências totais e trimadas, segundo as metodologias **M1 (total)** e **M2 (trimados)**. Ambas as árvores apresentaram topologias idênticas, porém a árvore que utiliza as sequências trimadas mostrou valores de *bootstrap* mais altos.

A figura 4.11 representa a mesma árvore na representação topológica de árvore de radiação para uma melhor apreciação da distribuição dos clados.

A fgura 4.12 mostra a comparação da árvore M2-i construída com o programa PHYML 2.4.4, modelo JTT, algoritmo de procura intensiva e a árvore M2-ii construída com programa PHYML 3, modelo RtREV, algoritmo SBR. Ambas as árvores apresentaram topologias muito similares, porém a construção da árvore com o programa PHYML 3, algoritmo SBR, demorou 28 dias comparados aos 2 dias requeridos para a construção da árvore com o programa PHYML 2.4.4, algoritmo de procura intensiva.

Ambas as árvores apresentaram três principais clados, os quais estão correlacionados com a presença dos ortólogos universais nas espécies analisadas. Os três clados obtidos correspondem aos três grupos de dados:

(i) 26 espécies que apresentam pelo menos 80% (25/31) dos ortólogos universais em seus genomas chamado grupo G1,

(ii) 12 espécies que apresentam entre 50-79% (15-24/31) dos ortólogos universais chamado grupo G2 e

(iii) 36 espécies que apresentam menos de 50% (1-14/31) dos ortólogos universais chamado grupo G3.

O grupo G1 foi formado por excavatas representados pelos tripanossomatídeos, tricomonas e diplomonas. Os Kinetoplastida foram caracterizados pela presença da monofilia formada pelos grupos parafiléticos *Leishmania* e *Trypanosoma*. *L. major* foi encontrada mais estreitamente relacionada a *L. infantum* que a *L. brasiliensis*, seguindo *L. enriettii* que pertence ao grupo G3. *T. brucei* e *T. cruzi* foram os representantes de *Trypanosoma*.

O grupo G2 foi formado por quatro grupos: (i) **Rodophyta**: *Porphyra, Gracilaria.* (ii) **Cryptophyta**: *Guillardia* e *Rhodomonas*; **Glaucocystophyceae**: *Cyanophora* e **Haptophyceae**: *Emiliania*. (A cryptophyta *Hemiselmis* agrupou próximo à Apicomplexa *Theileria* e *Babesia*). (iii) **Stramenopila**: *Odontella, Phaeodactylum* e *Heterosigma*. (iv) R**rodophyta**: *Cyanidioschyzon* e *Cyanidium.*

O grupo G3 foi formado pelos protozoários **Euglenozoa/Kinetoplastida**: *L. amazonensis*, *L. donovani*, *L. enriettii*; a **Alveolata/Apicomplexa**: *T. gondii* e *E. tenella*; a **Amebozoa**: *A. castellanii*, *A. polyphaga*, *A. healyi*, *D. citrinum*, *D. fasciculatum* e *Polysphondylium pallidum*; a **Alveolata/Ciliophora**: *P. aurelia*, *P. caudatum*, *T. malaccensis, T. pigmentosa, T. pyriformis* e *T. paravorax* e a **Euglenozoa/Euglenida**: *E. gracilis*, *E. longa.* Além disso, outros eucariontes foram encontrados: **Stramenopila**: *Cafeteria roenbergensis*, *Chrysodidymus synuroideus*, *Desmarestia viridis*, *Dictyota dichotoma*, *Fucus vesiculosus*, *Laminaria digitata, Ochromonas danica, Phytophthora infestans, P. sojae, P. ramorum, Pylaiella littoralis, Saprolegnia ferax* e *Thalassiosira pseudonana*; a **Rhodophyta**: *Chondrus crispus*, a **Malawimonadidae**: *Malawimonas jakobiformis*, a **Heterolobosea**: *Naegleria gruberi* e **Jakobidae** *Reclinomonas americana.*

D. discoideum pertencente ao grupo G1 encontrou-se mais próximo a *D. citrinum* do grupo G3. *D. fasciculatum* foi encontrado mais próximo a *P. pallidum*. No entanto, *L. amazonensis* e *L. donovani* não foram encontrados estreitamente relacionados. Os Amebozoa monofiléticos *A. castellanii* e *A. polyphaga* não estiveram filogeneticamente muito próximos, como esperado, com o grupo monofilético *A. healyi. E. gracilis* e *E. longa.*

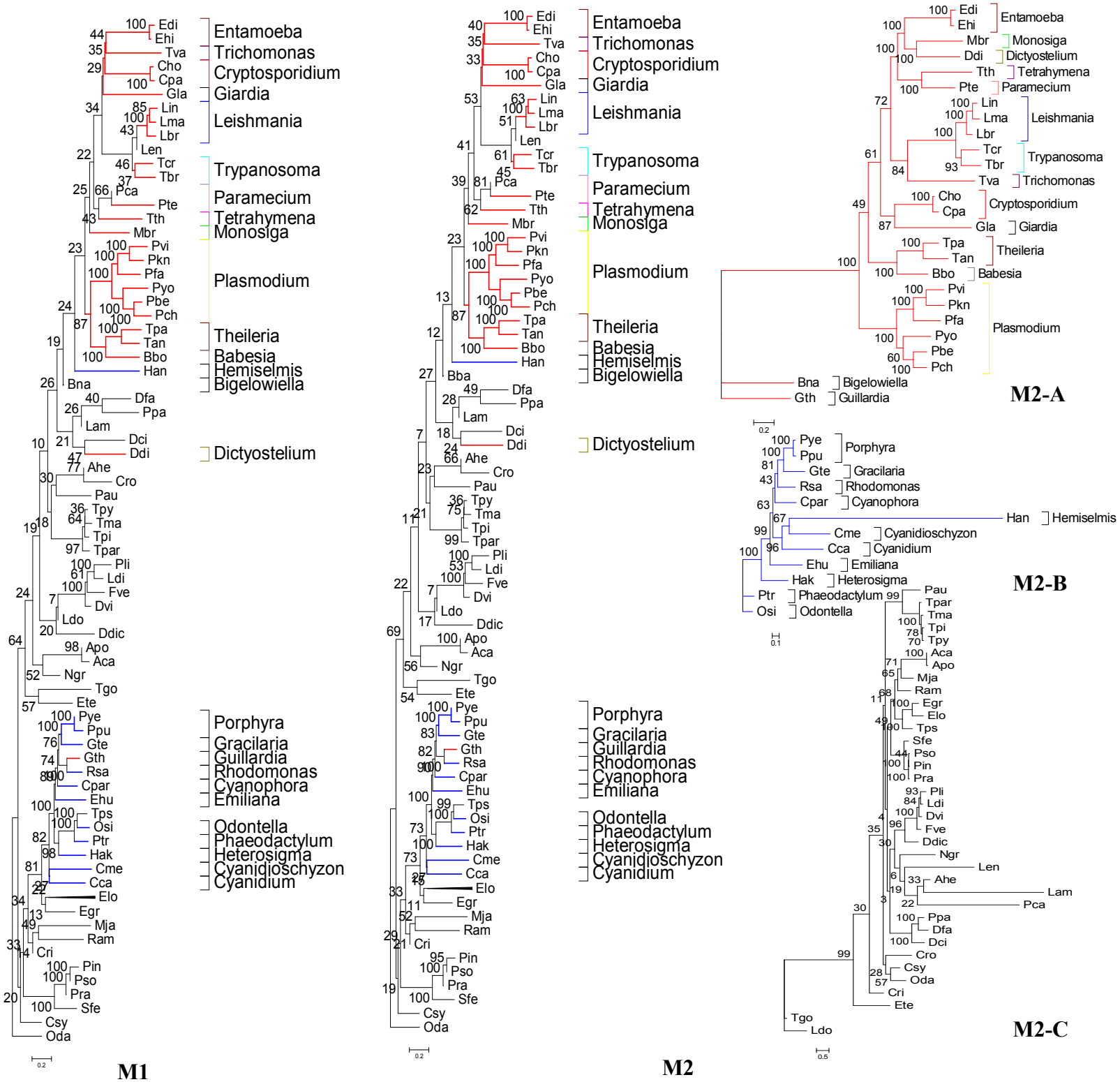


Figura 4.10 - Árvores filogenômicas da supermatriz de protozoários usando os alinhamentos de sequências M1 (totais) e M2 (trimados)

Foram usadas supermatrizes com 21.260 (M1) e 12.807 (M2) posições respectivamente de 74 espécies de protozoários. As árvores de máxima verossimilhança foram construídas com o PHYML, modelo evolutivo JTT e *bootstrap* 100. As cores diferenciam os três grupos: G1 vermelho, G2 azul e G3 preto.

A árvore M2-A pertencente ao grupo G1 foi criada usando o modelo evolutivo Blosum62, a M2-B do grupo G2 com o RtREV e a M3-B do grupo G3 com o WAG.

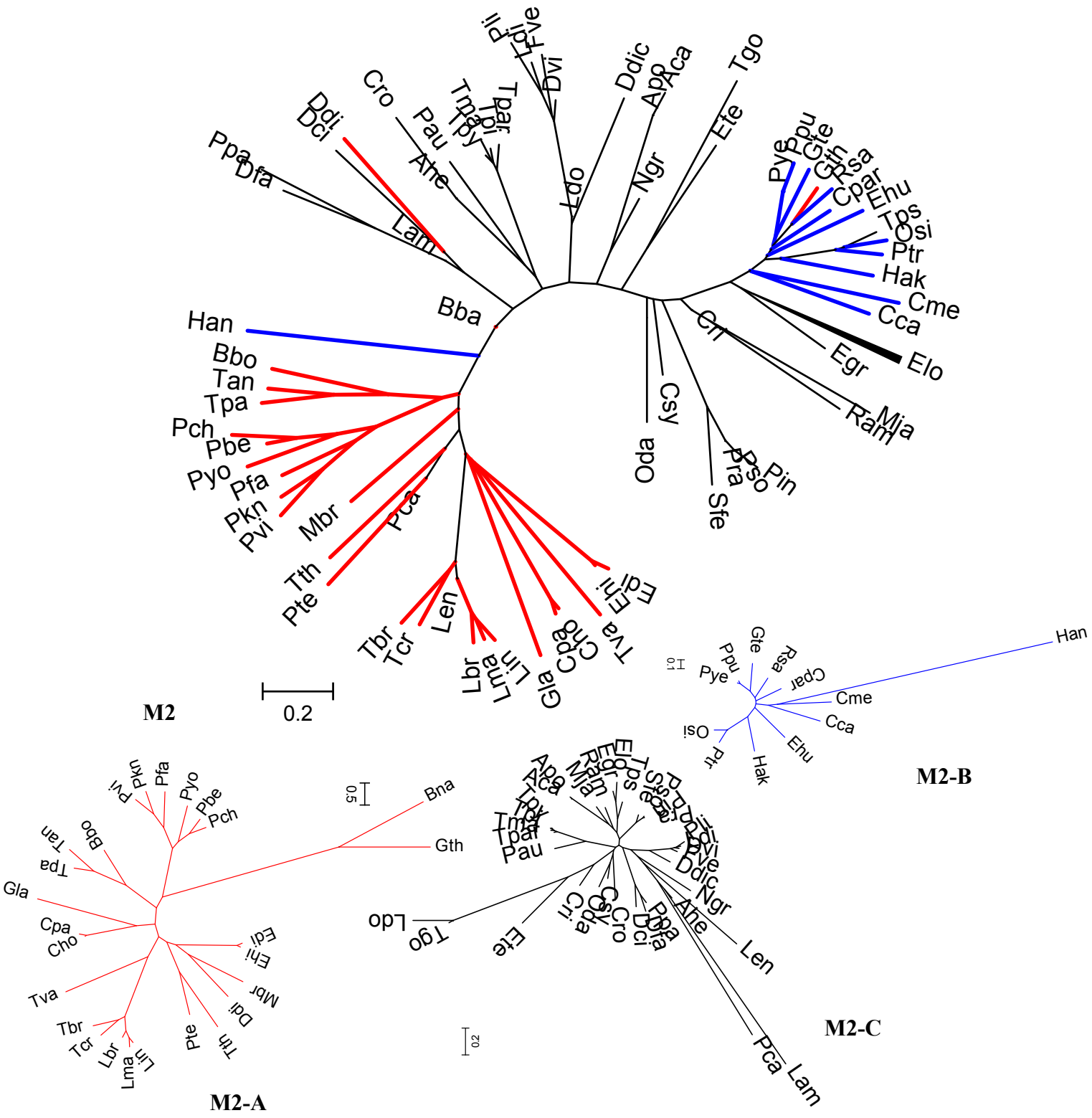


Figura 4.11 - Árvore (radiação) filogenômica da supermatriz de protozoários usando alinhamento de sequências M2 (trimadas)

Foi usada uma supermatriz com 12.807 posições (M2) de 74 espécies de protozoários. A árvore de máxima verossimilhança foi construída com o PHYML, modelo evolutivo JTT e *bootstrap* 100. As cores diferenciam os três grupos: o grupo G1 (M2-A) vermelho, o grupo G2 azul (M2-B) e o grupo G3 preto (M2-C).

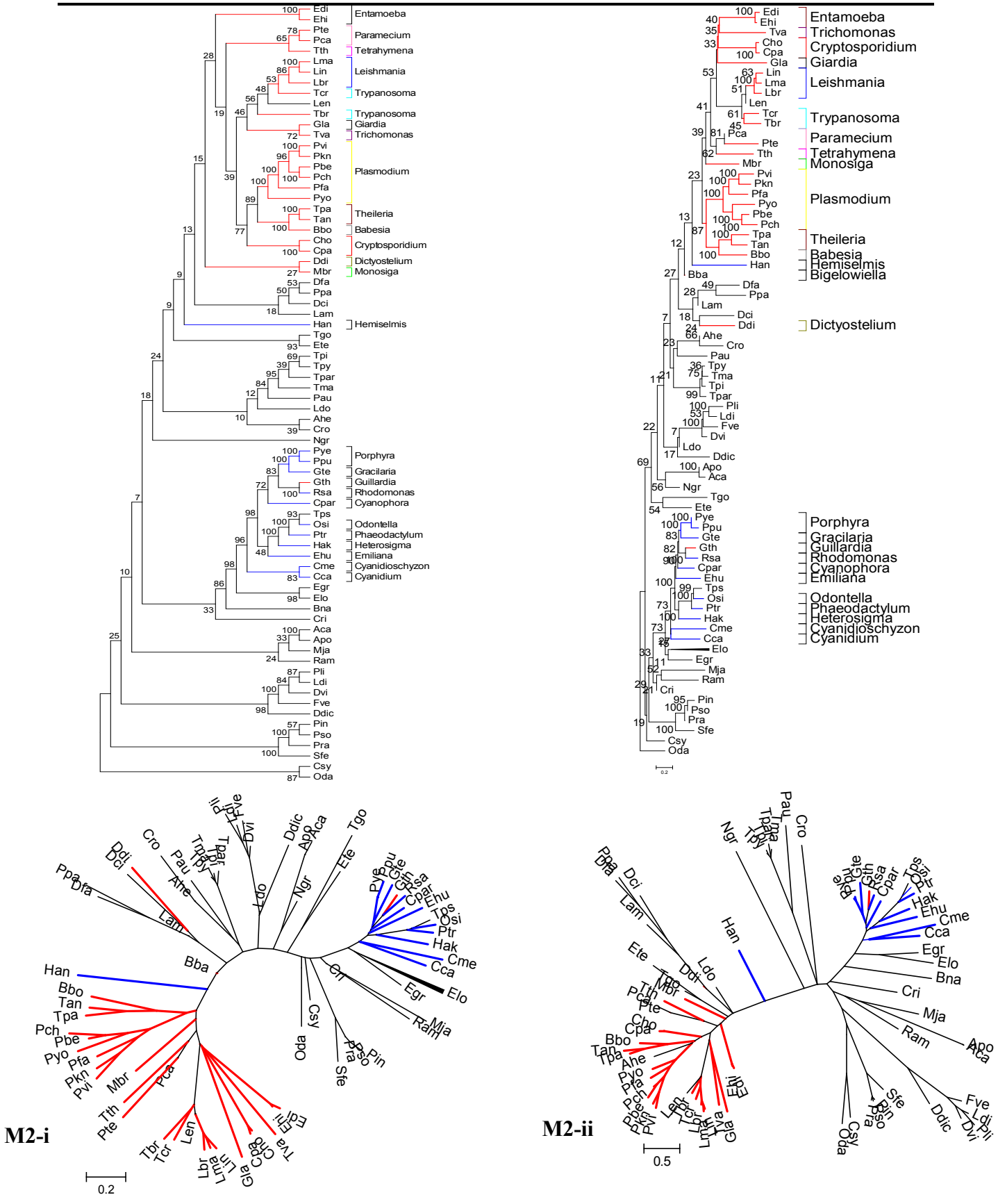


Figura 4.12 - Árvores (radiação) filogenômicas da supermatriz M2 dos modelos JTT e Blosum62

Foi usada uma supermatriz com 12.807 posições (M2) de 74 espécies de protozoários. M2-A foi construída com o modelo JTT, do programa PHYML 2.4.4, algoritmo de procura intensiva. M2-B foi construída com o modelo RtREV, do programa PHYML 3, algoritmo SBR As cores diferenciam os três grupos: o grupo G1 vermelho, o grupo G2 azul e o grupo G3 preto

4.3 Filogenômica dos EGM em protozoários

4.3.1 Análises filogenéticas

As árvores filogenéticas obtidas para os genes transcriptase reversa (figura 4.12, anexo 8.15), proteína gag (figura 4.13, anexo 8.16), proteína gag-pol (figura 4.14, anexo 8.17), proteína pol (figura 4.15, anexo 8.18), integrase, (figura 4.16, anexo 8.19) e ribonuclease H (figura 4.17, anexo 8.20) foram comparadas em termos de algoritmo, valores de *bootstrap* e consistência das relações filogenéticas e taxonômicas na topologia, apoiada na taxonomia atual do banco de dados TAXONOMY. As árvores mais consistentes foram obtidas com o algoritmo MV - executado com os programas RAXML e PHYML - e com o algoritmo IB - executado com o MRBAYES.

Os resultados da filogenia de cada gene construída com o RAXML foram representados por dos tipos de topologias nas seções **Mostrar só topologia** e **Topologia com comprimento de ramo,** respectivamente, contidas nas figuras 4.12-4.17:

(i) árvore que mostra a topologia com valores de *bootstrap* e

(ii) árvore que mostra os comprimentos de ramo.

As árvores construídas usando os outros programas filogenéticos: PHYML, MRBAYES (figura 4.12-4.17), PAUP-AV, PAUP-MP e WEIGHBOR (anexo 8.15-8.20) - não apresentaram diferenças significativas e foram também usadas para inferir filogenia e/ou esclarecer alguma incongruência. Os genes transcriptase reversa, ribonuclease H e integrase, estiveram formados por um elevado número de sequências pelo qual algumas das árvores filogenéticas não puderam ser construídas com o MRBAYES.

4.3.1.1 Análise filogenética da transcriptase reversa em protozoários

A árvore da transcriptase reversa foi dividida nos seguintes quatro grupos (figura 4.12):

(i) Retrotransposons não-LTR

As transcriptases reversas da ***G. lamblia*** (círculo preto) formaram um grupo monofilético e sofreram uma duplicação que as dividiu em dois subgrupos, conformados por:

(1) **as transcriptases reversas pertencentes aos retrotransposons não-LTR e não-LTR *like*** e

(2) **as transcriptases reversas incluídas** consideradas como não-LTR.

Esta última classificação não-LTR foi proporcionada com base na filogenia da nossa árvore, devido a que nas anotações das sequências, só figuram anotadas como transcriptases reversas, sem outra informação sobre o tipo de retrotransposons ao qual pertencem.

(1) O subgrupo **as transcriptases reversas pertencentes aos retrotransposons não-LTR e não-LTR *like*** foi encontrado mais próximo aos retrotransposons não-LTR SLACS de *L. braziliensis*, *T. brucei* e *T. cruzi*. Uma duplicação dividiu este subgrupo em duas agrupações contendo as trancriptases reversas com as seguintes anotações e denominações:

(1a) 1 GilM, LINE, 3 não-LTR *like* endonuclease, 1 não-LTR *like* pol, 1 não-LTR *like*, 1 não-LTR *like* ORF e 2 não-LTR LINE e

(1b) 2 GilT endonuclease, 1 não-LTR *like* endonuclease e 1 não-LTR *like* ORF.

(2) O subgrupo **as transcriptases reversas incluídas** foi representado por: 11 transcriptase reversa-LTR *like* endonuclease, 3 não-LTR LINE *like*,

O grupo monofilético composto por ***Trypanosoma* e *Leishmania*** apresentou-se como o mais próximo ao grupo *G. lamblia* acima descrito. O grupo foi composto por *Trypanosoma* (círculo celeste) e *Leishmania* (círculo azul) as quais sofreram duplicação que as dividiu em 3 subgrupos:

(1) ***T. brucei***, que apresentou 5 não-LTR SLACS (2 putativos, 1 *like*),

(2) ***T. cruzi***, que apresentou 1 não-LTR *like* CZAR endonuclease e

(3) ***L. brasiliensis***, que apresentou 6 não-LTR SLACS (5 *like*).

Outros grupos monofiléticos não-LTR apresentaram-se isolados e independentes na árvore, sendo a maioria deles registrados como transcriptases reversas-*like*. Estes grupos monofiléticos não-LTR foram definidos como os **grupos intermediários (GI)** a continuação:

(1) GI 1, formado por ***T. vaginalis*** que apresentou 2 não-LTR *like*,

(2) GI 2, formado por ***P. yoelii yoelii*** que apresentou 2 não-LTR *like* e ***P. infestans*** que apresentou 2 não-LTR *like*,

(3) GI 3, formado por ***T. brucei*** que apresentou 2 não-LTR INGI e 1 não-LTR *like*, e ***T. cruzi*** que apresentou 1 não-LTR *like* L1Tc e 1 não-LTR *like*.

(4) GI 4 foi formado por ***E. histolytica*** que apresentou 2 por TR não-LTR *like*,

(5) GI 5 foi formado por ***T. cruzi*** que apresentou 1 TR não-LTR *like* VIPER e

(6) GI 6 foi formado por ***T. thermophila*** que apresentou 9 TR não-LTR (6 *like* e 1 ORF).

Não foram encontradas transcriptases reversas pertencentes aos retrotransposons não-LTR em ***L. major***.

(ii) Retrotransposons LTR

Os retroelementos LTR formaram uma monofilia e foram divididos nos seguintes grupos:

(1) Os elementos **gypsy**, encontrados nas seguintes espécies:

P. ramorum (1 gypsy Pr-2), ***P. sojae*** (1 gypsy Pr-2, 2 gypsy Ps-1A), ***R. salina*** (8 gypsy Ty3) e ***P. yezoensis*** (1 gypsy Ty3 *like*).

(2) Os elementos **Copia**, encontrados nas seguintes espécies:

P. infestans (11 copia Ty1), ***P. parasitica*** (3 copia Ty1) e ***P. yezoensis*** (1 copia Ty1).

E. tenella foi incluída neste grupo baseada na topologia da árvore filogenética construída, mas as suas transcriptases reversas não possuíram especificação sobre o tipo de retrotransposons LTR ao qual pertencem.

Foram encontradas três sequências isoladas: *P. ramorum*: LTR gypsy Pr-0, LTR copia PR-1 e *P. yezoensis*: 1 gypsy Ty3, as quais encontraram-se separadas do grupo monofilético dos LTR acima descrito.

(iii) Telomerase

A telomerase formou os seguintes grupos monofiléticos:

(1) O grupo mais representativo encontrado nas espécies da família ***Trypanosomatidae***:

- 3 ***T. brucei*** (1 putativo), 4 ***T. cruzi*** (2 putativos),

- 2 ***L. infantum*** (1 putativo), 1 ***L. donovani***, 2 ***L. amazonensis***, 3 ***L. brasiliensis*** (2 putativos) e interessantemente 3 ***L. major*** (2 putativos).

(2) Outros grupos monofiléticos independentes foram detectados em:

1 ***P. tetraurelia***, 2 ***T. thermophila***, ***Cryptosporidium*** (2 ***C. hominis*** e 3 ***C. parvum)***, 3 ***G. lamblia***, 1 ***P. yoelii yoelii*** (putativo) e 2 ***T. annulata*** (2 putativos).

(iv) Íntron do Grupo II

O Íntron do Grupo II encontrou-se espalhado formando os seguintes grupos monofiléticos bem resolvidos: 18 ***P. littoralis***, 6 ***P. purpurea***, 1 ***R. sp***, 4 ***R. salina*** e 1 ***H. andersenii***.

P. littoralis foi o grupo mais diversificado e distribuído, formou três grupos principais os quais estiveram mais relacionados à *R. salina*. *P. purpurea* foi posicionada então considerada como um grupo intermediário entre eles. Foi reportada uma só sequência de *H. andersenii*.

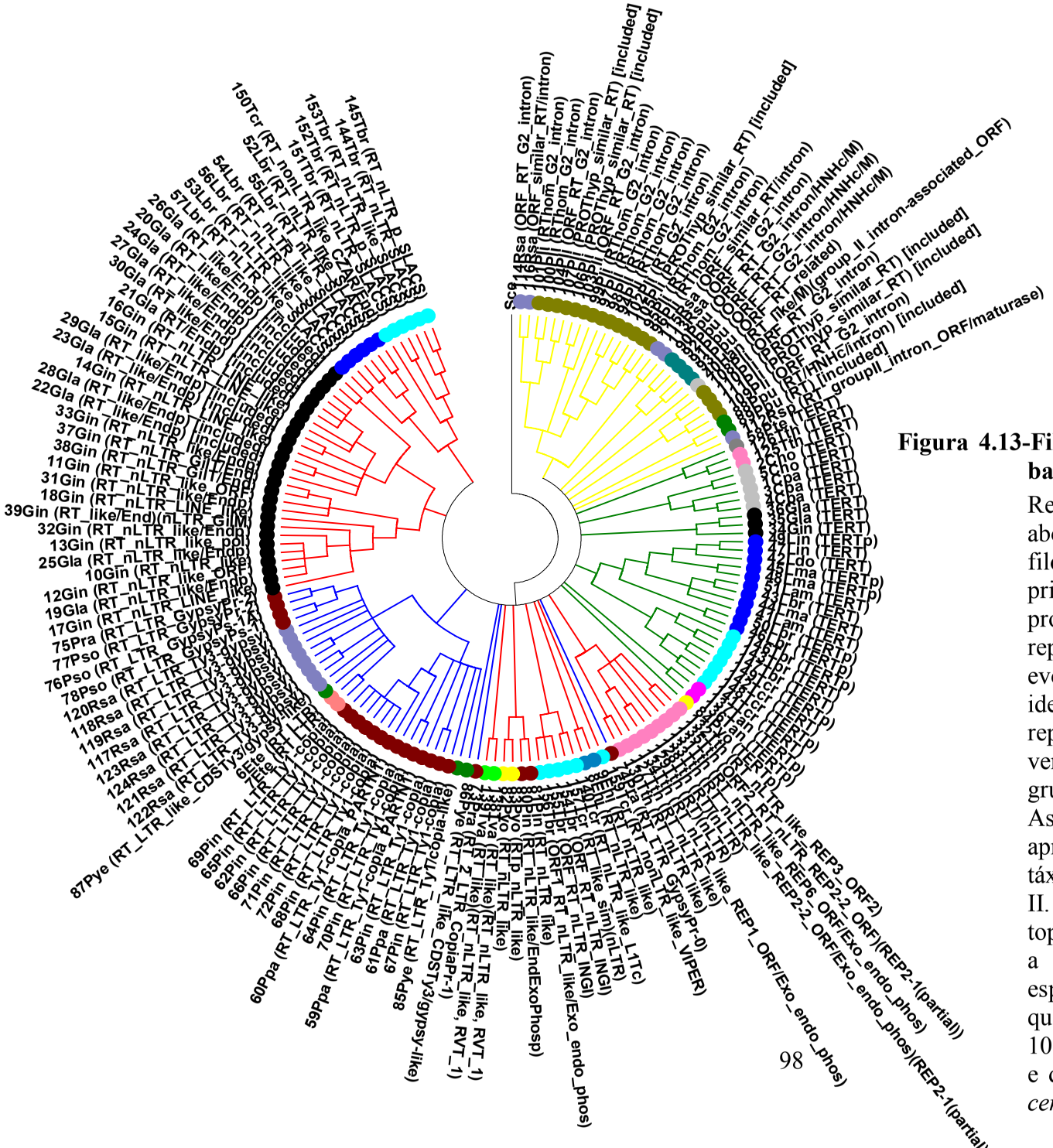


Figura 4.13-Filogenia dos retrotransposons em protozoários, baseada na transcriptase reversa

Representação esquemática mostrando uma abordagem filogenômica para os EGM baseados na filogenia da transcriptase reversa, para resolver os principais ramos da árvore da vida dos EGM dos protozoários. Esta árvore tem como objetivo representar uma visão consenso sobre as relações evolutivas entre os quatro ramos principais, identificados na árvore de MV. Os ramos vermelhos representaram aos não-LTR, os azuis aos LTR, os verdes às telomerases e as amarelas aos íntrons do grupo II.

As principais incertezas nos ramos da árvore apresentaram-se como anotações inconclusivas dos táxons nos grupos dos não-LTR e íntrons do grupo II. Estes táxons foram rotulados, baseados na topologia da filogenia, como **incluídos**, devido a que a anotação das transcriptases reversas não especificou o tipo de EGM ao qual pertencem. Note que a árvore foi construída usando o RAXML com 100 de *bootstrap*, matriz do modelo evolutivo WAG e distribuição gama. A árvore foi enraizada com *S. cerevisiae*.

4.3.1.2 Análise filogenética da proteína gag em protozoários

As árvores filogenéticas da proteína gag foram restritas as seguintes espécies: ***D. discoideum***, ***P. tricornotum***, ***T. pseudonana*** e ***P. infestans***. As árvores reconstruídas com os algoritmos de AV, MP e IB mostraram as topologias mais consistentes.

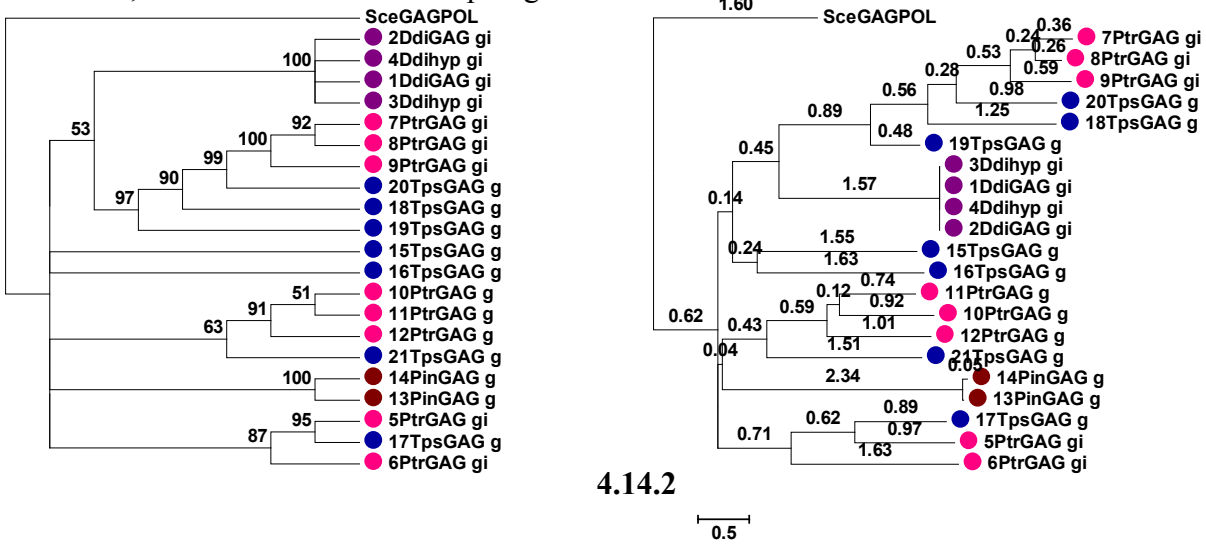


Figura 4.14- A filogenia da proteína gag

A árvore filogenética 4.14.1 mostra a topologia com valores de *bootstrap* e a 4.14.2 o comprimento dos ramos, respectivamente. As árvores foram construídas utilizando o RAXML, com 100 de *bootstrap*, matriz do modelo evolutivo BLOSUM62 e modelo de heterogeneidade de taxa de sítios invariáveis. A árvore foi enraizada com *S. cerevisiae*.

4.3.1.3 Análise filogenética da proteína gag-pol em protozoários

A proteína gag-pol foi encontrada formando dois grupos monofiléticos representativos em ***P. tricornotum*** e ***T. pseudonana*** como pertencendo aos retrotransposons Ty1/copia-*like*.

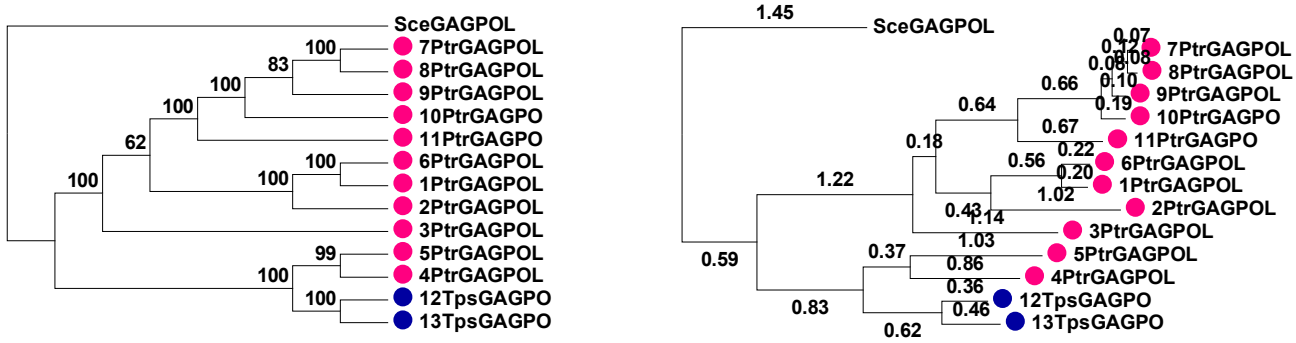
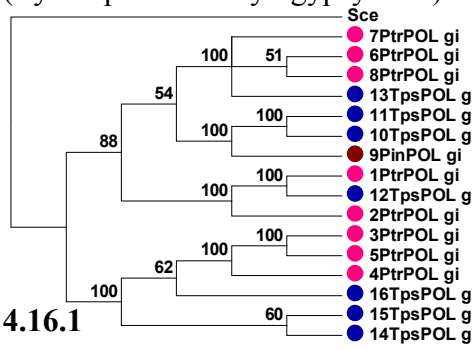


Figura 4.15-A filogenia da proteína gag-pol

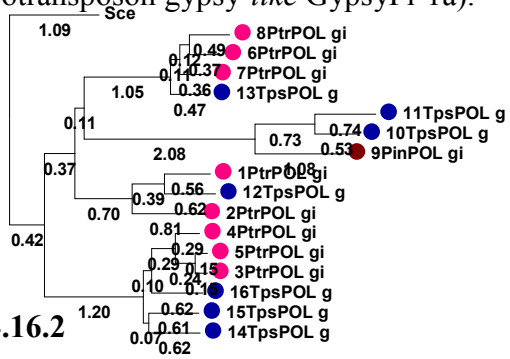
A árvore filogenética 4.15.1 mostra a topologia com valores de *bootstrap* e a 4.15.2 o comprimento dos ramos, respectivamente. As árvores foram construídas utilizando o RAXML, com 100 de *bootstrap*, matriz do modelo evolutivo WAG, distribuição gama e modelo de heterogeneidade de taxa de sítios invariáveis. A árvore foi enraizada com *S. cerevisiae*.

4.3.1.4 Análise filogenética da proteína pol em protozoários

A proteína pol foi encontrada em ***P. tricornotum*** (Ty1/copia-*like*), ***T. pseudonana*** (Ty1/copia-*like* e Ty3/gypsy-*like*) e ***P. infestans*** (retrotransposon gypsy-*like* GypsyPi-1a).



4.16.1



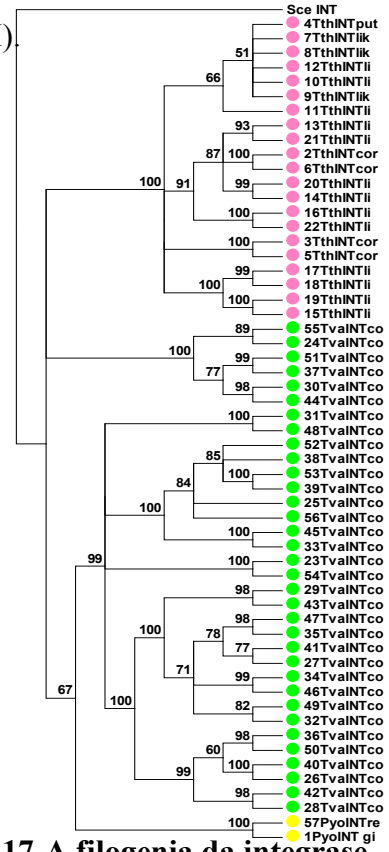
4.16.2

Figura 4.16.1-A filogenia da proteína pol

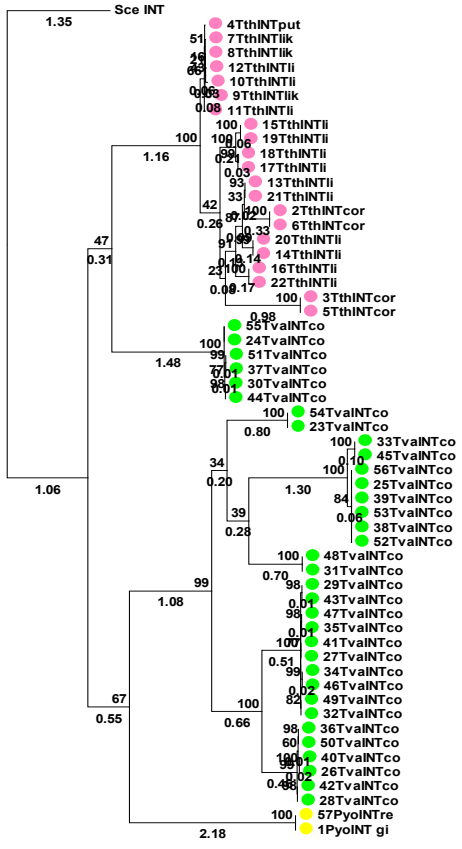
A árvore filogenética 4.16.1 mostra a topologia com valores de *bootstrap* e a 4.16.2 o comprimento dos ramos, respectivamente. As árvores foram construídas utilizando o RAXML, com 100 de *bootstrap*, matriz do modelo evolutivo BLOSUM62 e distribuição gama. A árvore foi enraizada com *S. cerevisiae*.

4.3.1.5 Análise filogenética da integrase em protozoários

A integrase formou três grupos monofiléticos nas espécies: ***T. thermophila*** (21 sequências), ***T. vaginalis*** (34 sequências) e ***P. yoelii yoelii*** (2 sequências, relacionadas a ribonuclease H).



4.17.1



4.17.2

Figura 4.17-A filogenia da integrase

A árvore filogenética 4.17.1 mostra a topologia com valores de *bootstrap* e a 4.17.2 o comprimento dos ramos, respectivamente. As árvores foram construídas utilizando o RAXML, com 100 de *bootstrap*, matriz do modelo evolutivo BLOSUM62 e distribuição gama. A árvore foi enraizada com *S. cerevisiae*.

4.3.1.6 Análise filogenética da ribonuclease H em protozoários

A ribonuclease H formou os seguintes grupos monofiléticos principais:

- A ribonuclease HI: ***Trypanosoma*** separou-se em subgrupos formados pelas espécies ***T. cruzi*** e ***T. brucei*** e ***Leishmania*** em ***L. major***, ***L. infantum*** e ***L. brasiliensis***.

- A ribonuclease H: grupo muito bem espalhado e esteve presente nas mesmas espécies e conformação que o grupo da ribonuclease HI, sugerindo uma possível paralogia com esta.

- Três sequências da ribonuclease: RHII: ***L. major*** (RHIIA, RHIIC) e ***T. brucei*** (RHII).

- Outras ribonucleases H: apresentando monofilias nas espécies ***B. bovis***, ***P. yoelii yoelii***, ***G. lamblia*** e ***T. thermophila***

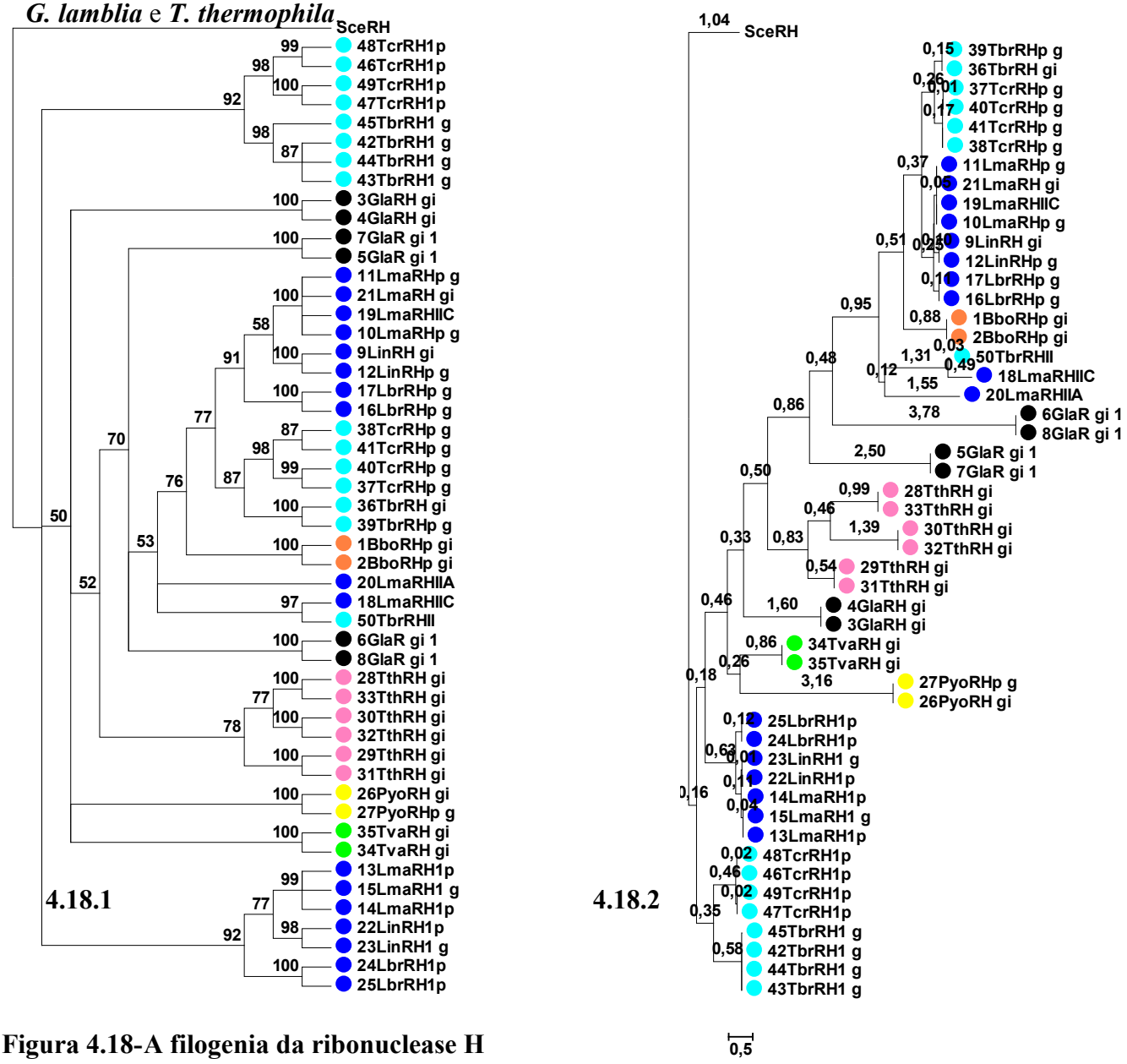


Figura 4.18-A filogenia da ribonuclease H

A árvore filogenética 4.18.1 mostra a topologia com valores de *bootstrap* e a 4.18.2 o comprimento dos ramos. As árvores foram construídas utilizando o RAXML, com 100 de *bootstrap*, matriz do modelo evolutivo BLOSUM62, distribuição gama e modelo de heterogeneidade de taxa de sítios invariáveis. A árvore foi enraizada com *S. cerevisiae*.

4.3.2 Estratégia para a detecção da seleção positiva de genes dos EGM em protozoários

O método de sítio por máxima verossimilhança permite a razão d_N/d_S (ω) variar entre os sítios (neste caso, cada códon é considerado um sítio) mas não entre ramos da árvore (NIELSEN; YANG, 1998). Foram utilizados três pares de modelos na análise de sítio. O primeiro par inclui os modelos M1 (aproximadamente neutro) e M2 (seleção positiva), o segundo par inclui M0 (razão única) e M3 (discreto) e o terceiro par inclui M7 (beta) e M8 (beta&ω). Para a análise de sítios usamos toda a região codificante dos genes da transcriptase reversa de todos os EGM encontrados (tabela 4.9), da transcriptase reversa dos retrotransposons não-LTR (tabela 4.10), da transcriptase reversa dos retrotransposons LTR (tabela 4.11), telomerase (tabela 4.12), proteína gag (tabela 4.13), proteína gag-pol (tabela 4.14), integrase (tabela 4.15) e ribonuclease H (tabela 4.16).

O modelo M1 (aproximadamente neutro) assume duas classes de sítio, com as proporções p_0 e p_1 estimadas a partir dos dados. O modelo M2 (seleção positiva) adiciona uma terceira classe com a razão ω_2 estimada. O modelo M7 (beta) assume uma distribuição beta para ω nos sítios. A distribuição beta é limitada ao intervalo (0, 1) e providencia uma hipótese nula flexível para testar a seleção positiva. Cálculos sugerem que esta distribuição se reduz ao modelo aproximadamente neutro (M1). O modelo M8 (beta&ω) adiciona uma outra classe de sítio ao modelo M7 (beta), com a razão ω estimada a partir dos dados.

4.3.2.1 Transcriptase reversa

O modelo M2 sugeriu que aproximadamente 7% dos sítios estão sob seleção positiva com ω_2 = 3,358 (tabela 4.9). Devido ao M2 ser uma extensão do M1, os dois modelos podem ser comparados usando um LRT. A estatística do teste LRT foi 2ΔlnL = 2 x ((-555909,48) - (-555937,61)) = 56,26 com P = 0,0003 e grau de liberdade (gl) = 2. Portanto, o modelo **M2** foi significantemente melhor que o modelo **M1**.

O modelo discreto (M3) com três classes (K=3) sugeriu que 8% dos sítios estão sob seleção positiva com ω_2 = 3,272 e identificou quatro aminoácidos sob seleção positiva. A estatística do teste LRT foi 2ΔlnL = 2 x ((-555909,43) - (-198839,80)) = 714139,26 com P < 0,0001 e gl=4. O modelo **M3** foi extrema e significativamente melhor que o modelo **M0**.

O modelo M8 sugeriu que aproximadamente 9% dos sítios estão sob seleção positiva com ω = 3,106 e identificou 4 sítios sob seleção positiva (os mesmos quatro sítios sugeridos por M3 e por M2, tabela 4.9). A estatística do teste LRT foi 2ΔlnL = 67,76 com P = 0,0002 e gl=2. O modelo **M8** foi significantemente melhor que o modelo **M7**.

De acordo com Yang e Nielsen (2002) a comparação M0-M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção positiva. Sendo assim, a pressão seletiva sob a transcriptase reversa parece ser muito variável ao longo de sua sequência. A distribuição de probabilidade posterior para os sítios da transcriptase reversa com os resultados do modelo M3 é apresentada na figura 4.18.

Tabela 4.9 - Resultados do teste de máxima verossimilhança para seleção positiva na transcriptase reversa

Modelo[1]	**p**[2]	l[3]	k[4]	**Parâmetros estimados**[5]		**LRT**[6]	**Sítios positivamente selecionados**[7]
M0	162	-198839,80	2.18	ω = 0,847		---	Nenhum
M1	163	-555937,61	2.31	p_0 = 0,464 p_1 = 0,535		---	Não permitido
M2	165	-555909,48	2.33	p_0 = 0,450 p_1 = 0,478 **p_2 = 0,071**	**ω_2 = 3,358**	&56,26 (P = 0,0003)	1110L** 1272N 1459A 1748Q
M3	166	-555909,43	2.33	p_0 = 0,391 p_1 = 0,532 **p_2 = 0,075**	ω_0 = 0,603 ω_1 = 0,954 **ω_2 = 3,272**	&&**714139,26** **(P < 0,0001)**	1110L** 1272N 1459A** 1748Q
M7	163	-555943,37	2.29	p = 3,33	q = 0,762	---	Não permitido
M8	165	-555909,49	2.33	p_0 = 0,913 p = 4,692 q = 1,296	**p_1 = 0,086** **ω = 3,106**	&67,76 (P = 0,0002)	**1110L** 1272N** **1459A** 1748Q****

[&] Valor estatisticamente significativo para a o teste LRT (P ≤ 0.05)
[&&] Valor extrema e estatisticamente significativo para a o teste LRT (P ≤ 0.0001)

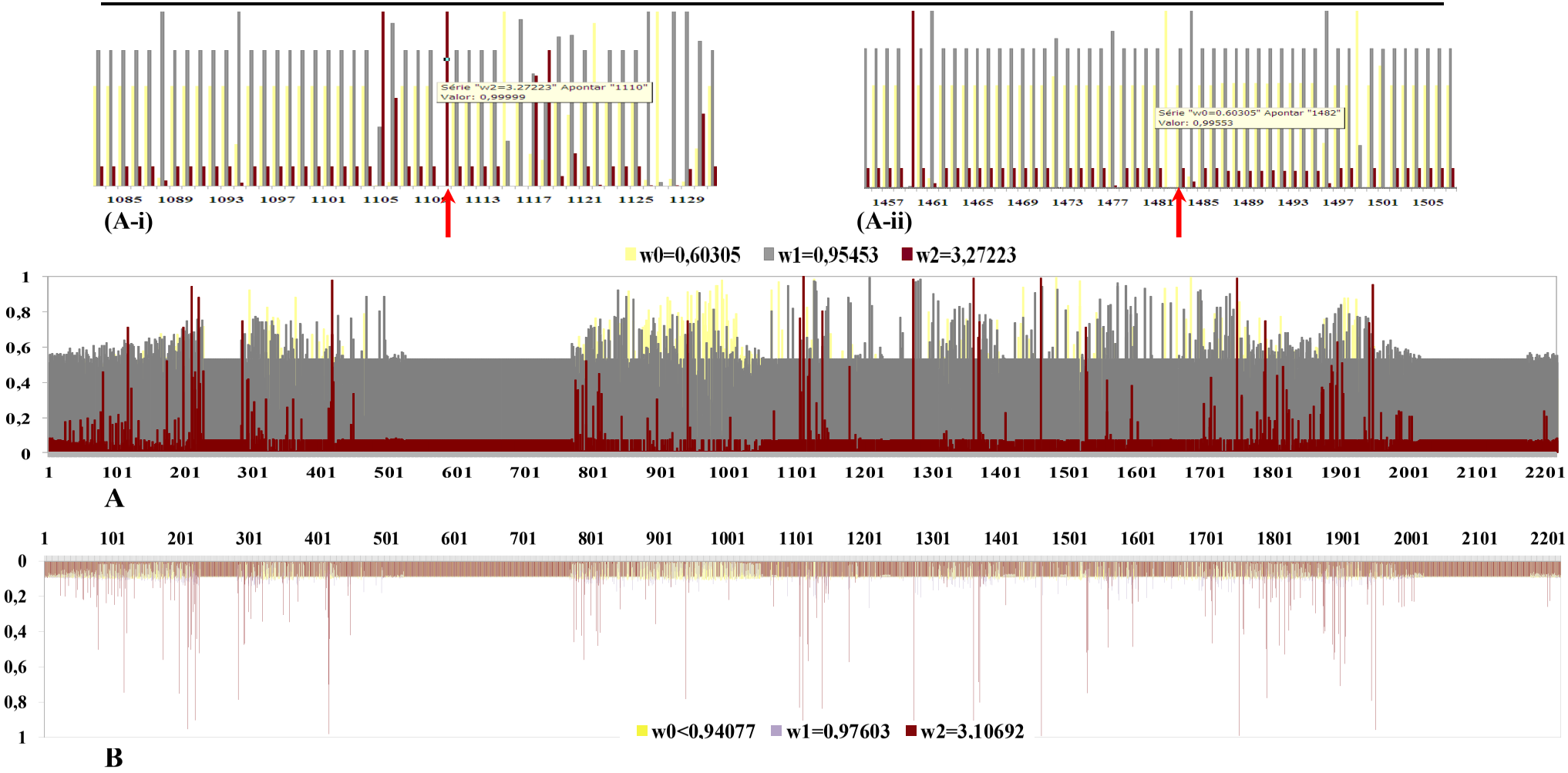


Figura 4.19-Probabilidade posterior das classes dos sítios da transcriptase reversa

(A) O modelo M3 (discreto) assume três classes de sítios na transcriptase reversa. As frequências estimadas e as razões ω para as três classes são $p_0 = 0,391$, $p_1 = 0,532$ e $p_2 = 0,075$ e $\omega_0 = 0,603$, $\omega_1 = 0,954$, $\omega_2 = 3,272$ (tabela 4.7). Essas *pk* são as distribuições a priori para cada sítio do códon (aminoácido). Os dados (configurações dos códons nas diferentes sequências) nos sítios mudam essa distribuição a priori para uma distribuição posterior, que pode ser muito diferente da a priori. Por exemplo, as probabilidades posteriores para as três classes no sítio 1482 são 0,9955, 0,0047 e 0,0000, e o sítio está sob seleção purificadora forte (A-i). Em contraste, as probabilidades posteriores no sítio 1110 são 0,0000, 0,0001 e 0,9999, e o sítio está certamente sob seleção de diversificação (A-ii).

(B) O modelo M8 (beta&ω) se ajusta aos dados com onze categorias. As razões ω para as onze categorias são 0.47940, 0.61393, 0.69081, 0.74801, 0.79501, 0.83584, 0.87273, 0.90722, 0.94077, 0.97603 e 3.10692 (tabela 4.7). As primeiras nove categorias foram combinadas no gráfico.

4.3.2.2 Retrotransposon não-LTR

O modelo M2 sugeriu que aproximadamente 7% dos sítios estão sob seleção positiva com ω_2 = 2,768 (tabela 4.10). A estatística do teste LRT foi 2ΔlnL = 2 x ((-291265,35) - (-291274,58)) = 18,46 com *P* = 0,0029 e grau de liberdade (gl) = 2. Portanto, o modelo **M2** foi significantemente melhor que o modelo **M1**.

O modelo M3 sugeriu que 1% dos sítios estão sob seleção positiva com ω_2 = 6,166 e identificou três aminoácidos sob seleção positiva. A estatística do teste LRT foi 2ΔlnL = 2 x ((-291269,80) - (-232547,60)) = 117444,4 com *P* < 0,0001 e gl=4. O modelo **M3** foi extrema e significativamente melhor que o modelo de razão única **M0**.

O modelo M8 sugeriu que aproximadamente 11% dos sítios estão sob seleção positiva com ω = 2,419, mas não identificou quais os sítios. A estatística do teste LRT foi com 2ΔlnL = 39,3 com *P* = 0,0006 e gl=2. O modelo **M8** foi significantemente melhor que o modelo **M7**.

De acordo com Yang e Nielsen (2002) a comparação M0-M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção positiva. Sendo assim, a pressão seletiva sob a transcriptase reversa dos retrotransposons não-LTR parece ser muito variável ao longo de sua sequência. A distribuição de probabilidade posterior para os sítios da transcriptase reversa dos retrotransposons não-LTR com os resultados do modelo M3 é apresentada na figura 4.19.

Tabela 4.10 - Resultados do teste de máxima verossimilhança para seleção positiva na transcriptase reversa dos retrotransposons não-LTR

Modelo[1]	p[2]	*l*[3]	*k*[4]	Parâmetros estimados[5]		LRT[6]	Sítios positivamente selecionados[7]
M0	87	-232547,60	1,74	ω = 0,688		---	Nenhum
M1	88	-291274,58	1,78	p_0 = 0,593 p_1 = 0,406		---	Não permitido
M2	90	-291265,35	1,78	p_0 = 0,680 p_1 = 0,247 **p_2 = 0,072**	**ω_2 = 2,768**	[&]18,46 (*P* = 0,0029)	Nenhum
M3	91	-291269,80	1,77	p_0 = 0,782 p_1 = 0,204 **p_2 = 0,013**	ω_0 = 0,567 ω_1 = 1,458 **ω_2 = 6,166**	[&&]**117444,4** **(*P* < 0,0001)**	737Q 874H 2091C
M7	88	-291284,57	1,77	p = 3,946 q = 1,661		---	Não permitido
M8	90	-291264,92	1,78	p_0 = 0,888 p = 10,017 q = 5,892	**p_1 = 0,111** **ω = 2,319**	[&]39,3 (*P* = 0,0006)	**Nenhum**

[&] Valor estatisticamente significativo para a o teste LRT (P ≤ 0.05)
[&&] Valor extrema e estatisticamente significativo para a o teste LRT (P ≤ 0.0001)

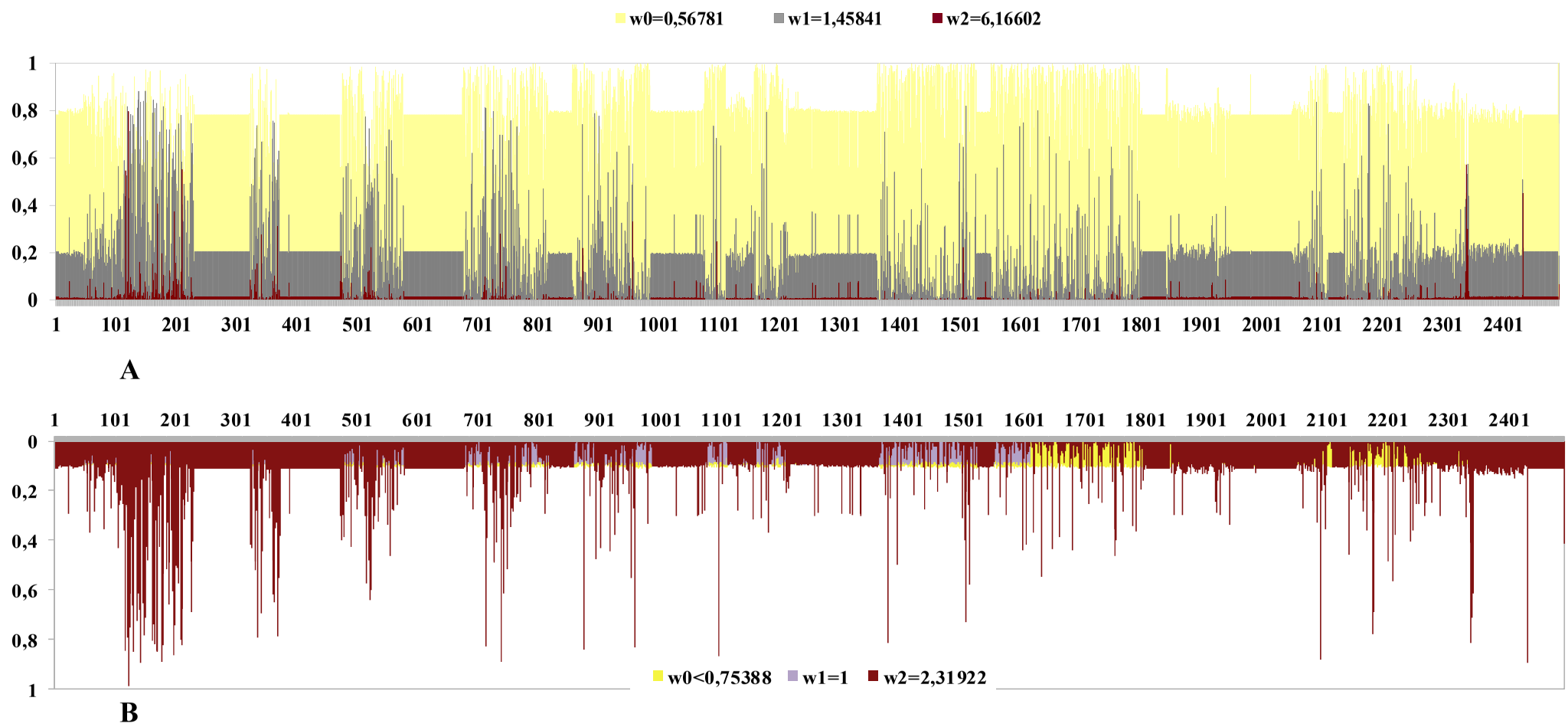


Figura 4.20-Probabilidade posterior das classes dos sítios da transcriptase reversa dos retrotransposons não-LTR

(A) O modelo M3 utilizado assume três classes de sítios no gene. As frequências estimadas e as razões ω para as três classes são p_0 = 0,782, p_1 = 0,204 e p_2 = 0,013 e ω_0 = 0,567, ω_1 = 1,458, ω_2 = 6,166 (tabela 4.8). Essas *pk* são as distribuições a priori para cada sítio do códon (aminoácido). Os dados nos sítios mudam essa distribuição a priori para uma distribuição posterior, que pode ser muito diferente da a priori. Por exemplo, as probabilidades posteriores para as três classes no sítio 713 são 0,9938, 0,0062 e 0,0000, e o sítio está sob seleção purificadora forte. Em contraste, as probabilidades posteriores no sítio 121 são 0,0012, 0,2025 e 0,79631, e o sítio está certamente sob seleção de diversificação.

(B) O modelo M8 (beta&ω) se ajusta aos dados com onze categorias. As razões ω para as onze categorias são 0.42681, 0.50413, 0.55049, 0.58723, 0.61972, 0.65050, 0.68143, 0.71460, 0.75388, 0.81338 e 2.31922 (tabela 4.8). As primeiras nove categorias foram combinadas no gráfico.

4.3.2.3 Retrotransposon LTR

O modelo M2 sugeriu que aproximadamente 36% dos sítios estão sob seleção positiva com ω_2 = 2,948 (tabela 4.11). A estatística do teste LRT é 2ΔlnL = 2 x ((-111526,66) - (-111572,40)) = 91,48 com P = 0,0001 e grau de liberdade (gl) = 2. Portanto, o modelo **M2** foi significantemente melhor que o modelo **M1** para os dados.

O modelo M3 sugeriu que 1% dos sítios estão sob seleção positiva com ω_2 = 289,208 e identificou que todos os aminoácidos estão sob seleção positiva. A estatística do teste LRT foi 2ΔlnL = 2 x ((-111515,22) - (-66462,44)) = 90105,56 com P < 0,0001 e gl=4. O modelo **M3** foi extrema e significativamente melhor que o modelo de razão única **M0**.

O modelo M8 sugiriu que aproximadamente 36% dos sítios estão sob seleção positiva com ω = 2,932 e identificou 8 sítios sob seleção positiva (os mesmos oito sítios sugeridos por M2 e sugeridos por M3, tabela 4.11). A estatística do teste LRT foi 2ΔlnL = 91,7 com P = 0,0001e gl=2.O modelo **M8** foi extrema e significantemente melhor que o modelo **M7**.

De acordo com Yang e Nielsen (2002) a comparação M0-M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção positiva. Sendo assim, a pressão seletiva sob a transcriptase reversa dos retrotransposons LTR parece ser muito variável ao longo de sua sequência. A distribuição de probabilidade posterior para os sítios da transcriptase reversa dos retrotransposons LTR com os resultados do modelo M3 é apresentada na figura 4.20.

Tabela 4.11 - Resultados do teste de máxima verossimilhança para seleção positiva na transcriptase reversa dos retrotransposons LTR

Modelo[1]	p[2]	l[3]	k[4]	Parâmetros estimados[5]		LRT[6]	Sítios positivamente selecionados[7]
M0	57	-66462,44	1,94	ω = 1,412		N/A	Nenhum
M1	58	-111572,40	1,85	p_0 = 0,047 p_1 = 0,952		N/A	Não permitido
M2	60	-111526,66	2,04	p_0 = 0,410 p_1 = 0,231 **p_2 = 0,357**	**ω_2 = 2,948**	[&]91,48, (P = 0,0001)	13T 18W 379T** 386P 820A 924D** 1153Q 1165T
M3	61	-111515,22	2,06	p_0 = 0,731 p_1 = 0,257 **p_2 = 0,011**	ω_0 = 1,229 ω_1 = 3,728 **ω_2 = 289,208**	**[&&]90105,56 (P < 0,0001)**	Todos **
M7	58	-111572,40	1,85	p = 51,657	q = 0,005	**N/A**	Não permitido
M8	60	-111526,55	2,03	p_0 = 0,641 p = 12,211 q = 0,005	**p_1 = 0,358** **ω = 2,932**	[&&]91,7 (P = 0,0001)	**13T 18W 379T** 386P 820A 924D** 1153Q 1165T**

[&] Valor estatisticamente significativo para a o teste LRT (P ≤ 0.05)
[&&] Valor extrema e estatisticamente significativo para a o teste LRT (P ≤ 0.0001)

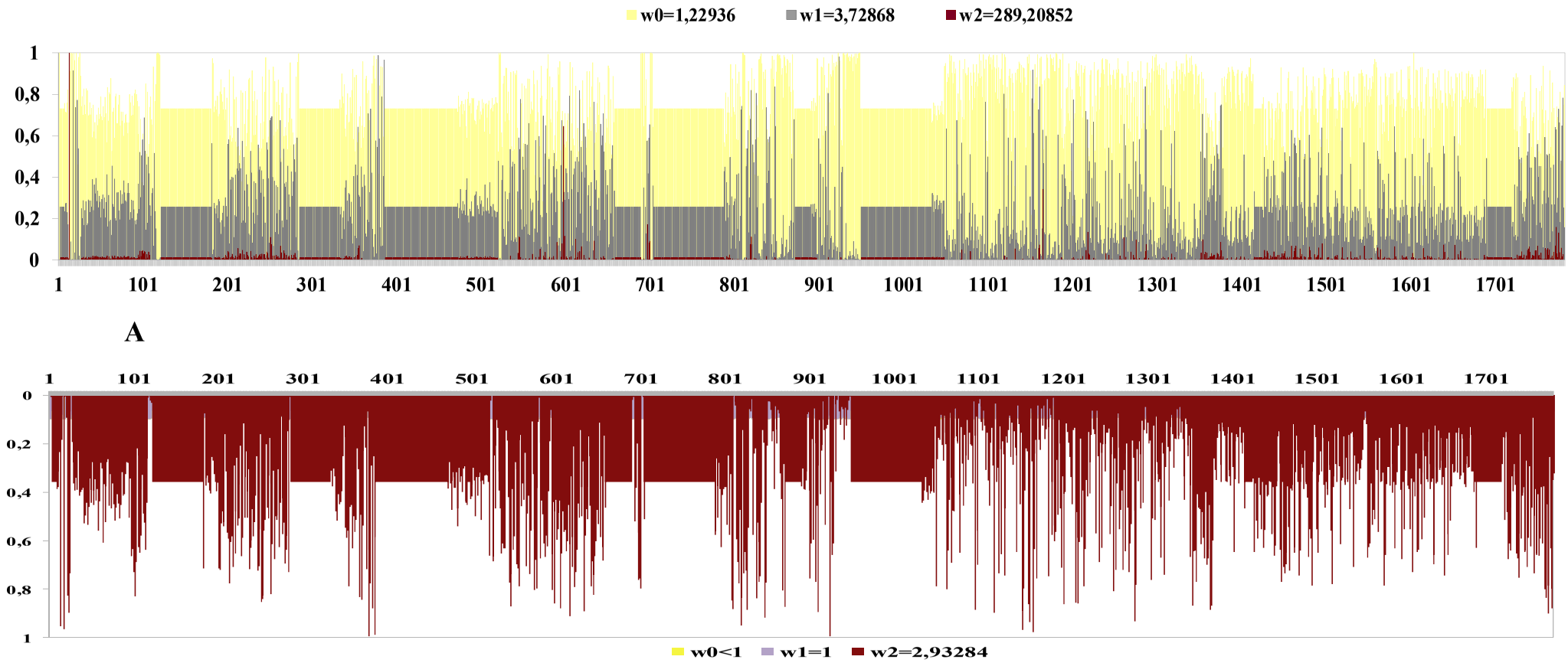


Figura 4.21-Probabilidade posterior das classes dos sítios da transcriptase reversa dos retrotransposons LTR

(A) O modelo M3 utilizado assume três classes de sítios no gene. As frequências estimadas e as razões ω para as três classes são p_0 = 0,731, p_1 = 0,257 e p_2 = 0,011 e ω_0 = 1,229, ω_1 = 3,728, ω_2 = 289,208 (tabela 4.9). Essas *pk* são as distribuições a priori para cada sítio do códon (aminoácido). Os dados (configurações dos códons nas diferentes sequências) nos sítios mudam essa distribuição a priori para uma distribuição posterior, que pode ser muito diferente da a priori. Por exemplo, as probabilidades posteriores para as três classes no sítio 19 são 0,9986, 0,0014 e 0,0000, e o sítio está sob seleção purificadora forte. Em contraste, as probabilidades posteriores no sítio 13 são 0,0000, 0,0004 e 0,9996, e o sítio está certamente sob seleção de diversificação.

(B) O modelo M8 (beta&ω) se ajusta aos dados com onze categorias. As razões ω para as onze categorias são 0.47940, 0.61393, 0.69081, 0.74801, 0.79501, 0.83584, 0.87273, 0.90722, 0.94077, 0.97603 e 3.10692 (tabela 4.9). As primeiras nove categorias foram combinadas no gráfico.

4.3.2.4 Telomerase

O modelo M2 sugeriu que aproximadamente 14% dos sítios estão sob seleção positiva com ω_2 = 1,000 (tabela 4.12). A estatística do teste LRT foi 2ΔlnL = 2 x ((-329108,71) - (-329118,95)) = 20,48 com P = 0,0024 e grau de liberdade (gl) = 2. Portanto, o modelo **M2** foi significantemente melhor que o modelo M1.

O modelo M3 sugeriu que 1% dos sítios estão sob seleção positiva com ω_2 = 8,257 e identificou aproximadamente 250 aminoácidos sob seleção positiva. A estatística do teste LRT foi 2ΔlnL = 2 x ((-329106,77) - (-268847,52)) = 120518,5 com P < 0,0001 e gl=4. O modelo **M3** foi extrema e significativamente melhor que o modelo de razão única **M0**.

O modelo M8 modelo sugeriu que aproximadamente 16% dos sítios estão sob seleção positiva com ω = 1,971 mas não especificou quais os sítios (tabela 4.12). A estatística do teste LRT foi 2ΔlnL = 20,74 com P = 0,0023e gl=2. O modelo **M8** foi extrema e significantemente melhor que o modelo **M7**.

De acordo com Yang e Nielsen (2002) a comparação M0-M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção positiva. Sendo assim, a pressão seletiva sob a telomerase parece ser muito variável ao longo de sua sequência. A distribuição de probabilidade posterior para os sítios da telomerase com os resultados do modelo M3 é apresentada na figura 4.21.

Tabela 4.12 - Resultados do teste de máxima verossimilhança para a seleção positiva na telomerase

Modelo[1]	p[2]	*l*[3]	*k*[4]	Parâmetros estimados[5]		LRT[6]	Sítios positivamente selecionados[7]
M0	47	-268847,52	2,91	ω = 1,040		N/A	Nenhum
M1	48	-329118,95	2,93	p_0 = 0,077 p_1 = 0,922		N/A	Não permitido
M2	50	-329108,71	3,02	p_0 = 0,067 p_1 = 0,796 **p_2 = 0,136**	**ω_2 = 1,00**	[&]20,48 (*P* = 0,0024)	Nenhum
M3	51	-329106,77	3,01	P_0 = 0,207 p_1 = 0,786 **p_2 = 0,006**	ω_0 = 0,006 ω_1 = 1,224 **ω_2 = 8,257**	[&&]**120518,5** **(*P* < 0,0001)**	3V 23E** 26E 42D** 45G** 47L 53E** 55L 58E 74R 77T 116L 117I 118L 120E 122G 124S 130S 136G** 140D** 141T 142R 143S 144E 145L 148L 176V 178K 188H 214S** 224P 231Q 259V** 260I 263Y 266F 274I 276F 287S 290F 291A** 292V 296E 307N 308K 314L 316T 317S 321L 324F 334F 335I 339I 340F 342N 351R 369K 372R 374Y 407S 409S 412I 413S** 417S 427R 432C** 443K 446R 454Y 498P 499K 544Y 559I 562D 565I** 579T 583K 627N 643I 688R 690V 693K 699I** 719K 720E 723Q 734K 738K 739L** 740I 744S** 748E** 754N 755K 757G 763F** 765T 778S 785R** 787H 795W 806W 808S** 811K 818L 819L 832R 843Y 846K 852F 876Y 877R 883R 887T 944Y 951R 952L 954R 955I 961E 988I 992K 999F 1003N 1017F 1018L 1025L 1066L 1070K 1071E 1072N 1079I 1082K 1089K 1090V 1092K 1096K 1108W 1116S 1120L 1128A** 1130I 1136S 1137R 1166V 1187K 1193F 1205S 1213Y 1217R** 1220S 1254T** 1258S** 1300I 1328N 1329P** 1330C 1331Q* 1332S 1333Q 1348K 1354G 1359F 1361H** 1367R 1374T 1394L 1408S 1411R 1414M 1415F 1416Y 1419R** 1423P** 1426H 1429P 1433P 1439R 1445L 1461R 1467K** 1470R 1474M 1475K 1476K 1491N 1495G 1500F 1504T** 1505A 1506I 1508I 1510W 1520Y 1527P 1539Q 1551K 1553F 1558I 1559C 1570T 1572I 1577E 1591G** 1593F 1601P 1610I
M7	48	-329119,16	2,93	p = 0,138 q = 0,007		N/A	Não permitido
M8	50	-329108,79	3,02	p_0 = 0,841 p = 0,219 q = 0,012	**p_1 = 0,158** **ω = 1,971**	[&]20,74 (*P* = 0,0023)	**Nenhum**

[&] Valor estatisticamente significativo para a o teste LRT (P ≤ 0.05)
[&&] Valor extrema e estatisticamente significativo para a o teste LRT (P ≤ 0.0001)

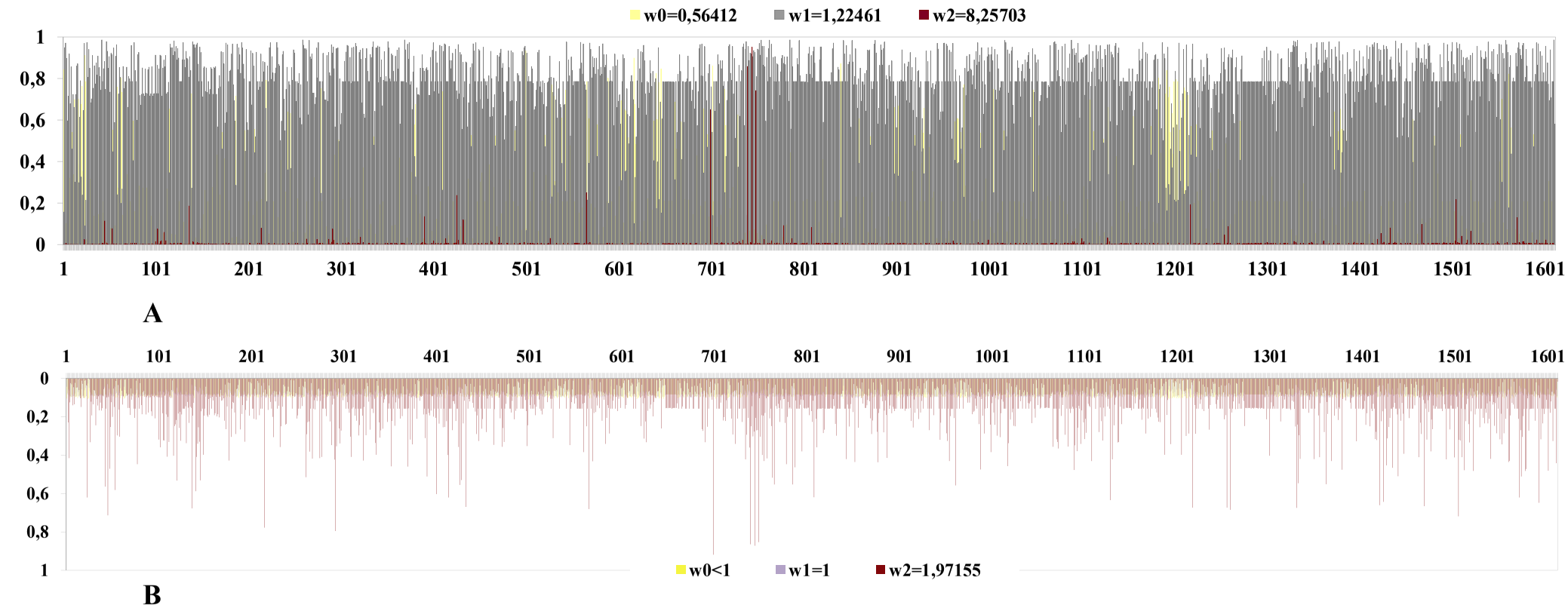


Figura 4.22-Probabilidade posterior das classes dos sítios da telomerase

(A) O modelo M3 utilizado assume três classes de sítios no gene. As frequências estimadas e as razões ω para as três classes são p_0 = 0,207, p_1 = 0,786 e p_2 = 0,006 e ω_0 = 0,006, ω_1 = 1,224, ω_2 = 8,257 (tabela 4.10). Essas *pk* são as distribuições a priori para cada sítio do códon (aminoácido). Os dados (configurações dos códons nas diferentes sequências) nos sítios mudam essa distribuição a priori para uma distribuição posterior, que pode ser muito diferente da a priori. Por exemplo, as probabilidades posteriores para as três classes no sítio 500 são 0,9323, 0,0677 e 0,0000, e o sítio está sob seleção purificadora forte. Em contraste, as probabilidades posteriores no sítio 744 são 0,0001, 0,0489 e 0,951, e o sítio está certamente sob seleção de diversificação.

(B) O modelo M8 (beta&ω) se ajusta aos dados com onze categorias. As razões ω para as onze categorias são 0.43559, 0.99985, 1.00000, 1.00000, 1.00000, 1.00000, 1.00000, 1.00000, 1.00000, 1.00000 e 1.97155 (tabela 4.10). As primeiras nove categorias foram combinadas no gráfico.

4.3.2.5 Proteína gag

O modelo M2 sugeriu que aproximadamente 22% dos sítios estão sob seleção positiva com ω_2 = 2,498 (tabela 4.13). A estatística do teste LRT foi 2ΔlnL = 2 x ((-157983,71) - (-157993,01)) = 18,62 com P = 0,0029 e grau de liberdade (gl) = 2. Portanto, o modelo **M2** foi significantemente melhor que o modelo **M1**.

O modelo M3 sugeriu que 46% dos sítios estão sob seleção positiva com ω_2 = 1,344 e identificou seis aminoácidos sob seleção positiva. A estatística do teste LRT foi 2ΔlnL = 2 x ((-157974,39) - (-157983,71)) = 189375,66 com P < 0,0001 e gl=4. O modelo **M3** foi extrema e significativamente melhor que o modelo de razão única **M0**.

O modelo M8 sugeriu que aproximadamente 94% dos sítios estão sob seleção positiva com ω = 1,000, mas não identificou quais os sítios (tabela 4.13). O modelo **M8** foi significantemente melhor que o modelo **M7** com 2ΔlnL = 7,02 com P = 0,0197 e gl=2.

De acordo com Yang e Nielsen (2002) a comparação M0-M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção positiva. Sendo assim, a pressão seletiva sob a proteína gag parece ser muito variável ao longo de sua sequência. A distribuição de probabilidade posterior para os sítios da proteína gag com os resultados do modelo M3 é apresentada na figura 4.22.

Tabela 4.13 - Resultados do teste de máxima verossimilhança para a seleção positiva na proteína gag

Modelo[1]	p[2]	l[3]	k[4]	Parâmetros estimados[5]		LRT[6]	Sítios positivamente selecionados[7]
M0	43	-63286,56	2,00	ω = 0,850		N/A	Nenhum
M1	44	-157993,02	2,08	p_0 = 0,002 p_1 = 0,997		N/A	Não permitido
M2	46	-157983,71	2,25	p_0 = 0,002 p_1 = 0,781 **p_2 = 0,215**	**ω_2 = 2,498**	&18,62 (P = 0,0029)	Nenhum
M3	47	-157974,39	2,16	p_0 = 0,002 p_1 = 0,535 **p_2 = 0,462**	ω_0 = 0,036 ω_1 = 0,722 **ω_2 = 1,344**	&&**189375,66 (P < 0,0001)**	102Y 1015M 1471T 1522L 1526L 1538Q
M7	44	-157992,66	2,07	p = 2,066 q = 0,295		N/A	Não permitido
M8	46	-157989,15	2,09	p_0 = 0,057 p = 1,066 q = 0,686	**p_1 = 0,942** **ω = 1,000**	&7,02 (P = 0,0197)	**Nenhum**

[&] Valor estatisticamente significativo para a o teste LRT (P ≤ 0.05)
[&&] Valor extrema e estatisticamente significativo para a o teste LRT (P ≤ 0.0001)

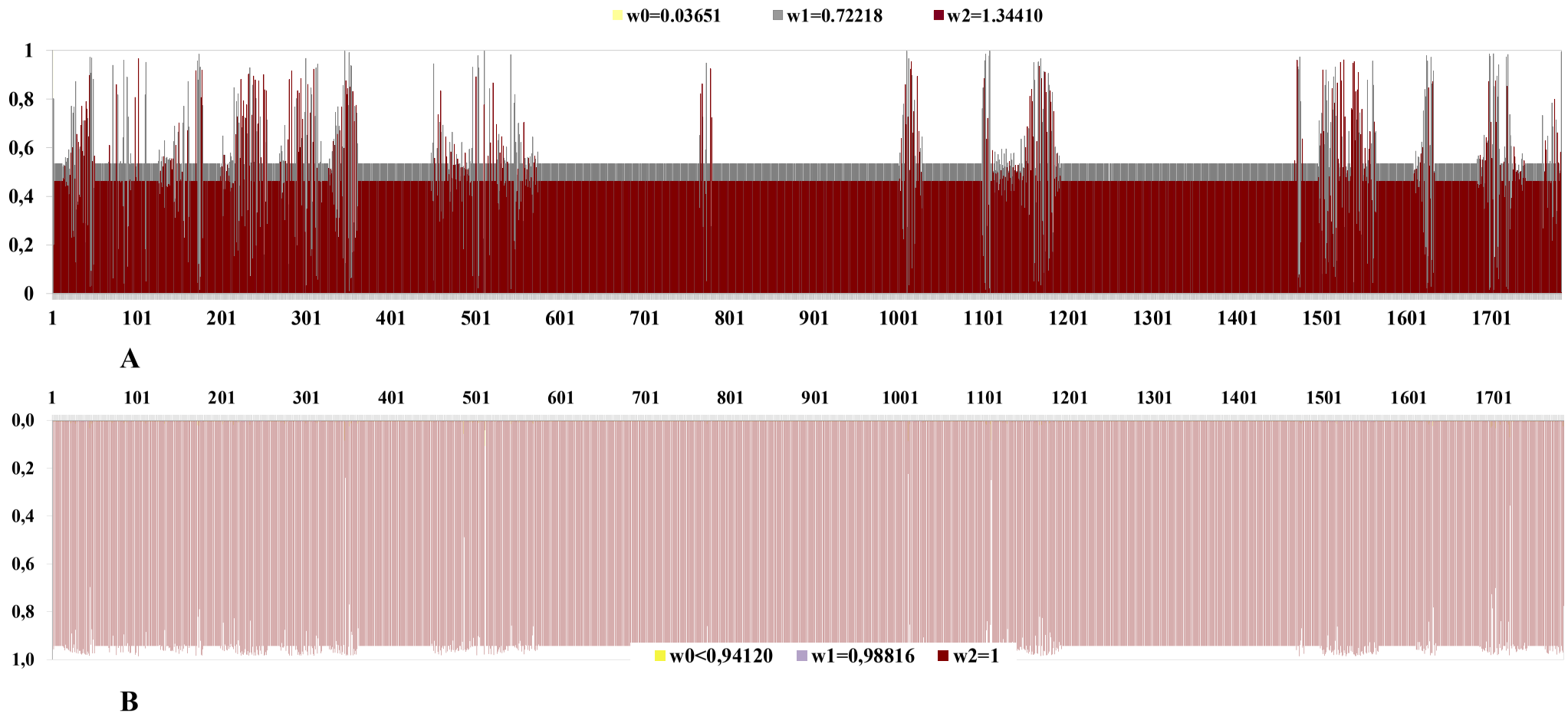


Figura 4.23-Probabilidade posterior das classes dos sítios da proteína gag

(A) O modelo M3 utilizado assume três classes de sítios no gene. As frequências estimadas e as razões ω para as três classes são p_0 = 0,002, p_1 = 0,535 e p_2 = 0,462 e ω_0 = 0,036, ω_1 = 0,722, ω_2 = 1,344 (tabela 4.11). Essas *pk* são as distribuições a priori para cada sítio do códon (aminoácido). Os dados (configurações dos códons nas diferentes sequências) nos sítios mudam essa distribuição a priori para uma distribuição posterior, que pode ser muito diferente da a priori. Por exemplo, as probabilidades posteriores no sítio 102 são 0,0000, 0,0336 e 0,9663, e o sítio está certamente sob seleção de diversificação.
(B) O modelo M8 (beta&ω) se ajusta aos dados com onze categorias. As razões ω para as onze categorias são 0.08577, 0.23423, 0.36835, 0.49120, 0.60360, 0.70560, 0.79672, 0.87592, 0.94120, 0.98816 e 1.00000 (tabela 4.11). As primeiras nove categorias foram combinadas no gráfico.

4.3.2.6 Proteína gag-pol

O modelo M2 sugeriu que aproximadamente 29% dos sítios estão sob seleção positiva com ω_2 = 1.893 (tabela 4.14). A estatística do teste LRT foi 2ΔlnL = 2 x ((-434525,60) - (-434576,05)) = 100,9 com P < 0,0001 e grau de liberdade (gl) = 2. Portanto, o modelo **M2** foi extrema e significantemente melhor que o modelo **M1**.

O modelo M3 sugeriu que 51% dos sítios estão sob seleção positiva com ω_2 = 1,564 e identificou 107 aminoácidos sob seleção positiva com p < 0,001. A estatística do teste LRT foi 2ΔlnL = 2 x ((-434522,28) - (-160720,60)) = 547603,36 com P < 0,0001 e gl=4. O modelo **M3** foi extrema e significativamente melhor que o modelo de razão única **M0**.

O modelo M8 sugeriu que aproximadamente 40% dos sítios estão sob seleção positiva com ω = 1,695 e identificou 13 sítios sob seleção positiva (sítios sugeridos também por M3, tabela 4.14). A estatística do teste LRT foi 2ΔlnL = 139,48 com P < 0,0001 e gl=2. O modelo **M8** foi extrema e significantemente melhor que o modelo **M7**.

De acordo com Yang e Nielsen (2002) a comparação M0-M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção positiva. Sendo assim, a pressão seletiva sob a proteína gag-pol parece ser muito variável ao longo de sua sequência. A distribuição de probabilidade posterior para os sítios da proteína gag-pol com os resultados do modelo M3 é apresentada na figura 4.23.

Tabela 4.14 - Resultados do teste de máxima verossimilhança para a seleção positiva na proteína gag-pol

Modelo[1]	p[2]	l[3]	k[4]	Parâmetros estimados[5]		LRT[6]	Sítios positivamente selecionados[7]
M0	27	-160720,60	2,00	$\omega = 0,939$		N/A	Nenhum
M1	28	-434576,05	2,05	$p_0 = 0,209$ $p_1 = 0,790$		N/A	Não permitido
M2	30	-434525,60	2,16	$p_0 = 0,002$ $p_1 = 0,781$ **$p_2 = 0,285$**	**$\omega_2 = 1,893$**	[&]100,9 ($P < 0,0001$)	261I
M3	31	-434522,28	2,16	$p_0 = 0,046$ $p_1 = 0,444$ **$p_2 = 0,508$**	$\omega_0 = 0,290$ $\omega_1 = 0,715$ **$\omega_2 = 1,564$**	[&&]**547603,36** ($P < 0,0001$)	35H 37N 96I 136K 145I 147R 160K 183S 186L 189T 191F 192S 195S 245T 261I** 288T** 289H 316H 319K 332R** 348F 402H 409L 410Q 413R 415F 421T 432A 440Y 446V 448S 451Y 454N 460T 488N 489R 496A 502L 504H 505L 526N 532L 534I 535L** 540N 543G 551N** 552Q** 567H 568H** 573M 579L 584K** 585N 605I 611R 612L 618I 622Q 661K** 664T 665F 682K 692T 722V 744L 832K 838E 841Q 843L 844H 845T 893C 930A 1037G 1111D 1112Q 1118L 1226H 1276W 1277T 1299Y 1323Y 1331K 1367Y 1375A 1380F 1397Y 1430R 1452S 1478T 1503E 1514S 1550L 1576N 1577I 1706Y 1721Q 1759I 1777T 1783K 1787H 1791N 1869M** 1870K 1983Q 1986M
M7	28	-434592,03	2,04	$p = 1,115$ $q = 0,164$		N/A	Não permitido
M8	30	-434522,29	2,16	$p_0 = 0,604$ $p = 2,268$ $q = 0,679$	**$p_1 = 0,395$** **$\omega = 1,695$**	[&]139,48 ($P < 0,0001$)	**261I 288T 332R 534I 535L 551N 552Q 568H 584K 661K 845T 1869M 1870K**

[&] Valor estatisticamente significativo para a o teste LRT ($P \leq 0.05$)
[&&] Valor extrema e estatisticamente significativo para a o teste LRT ($P \leq 0.0001$)

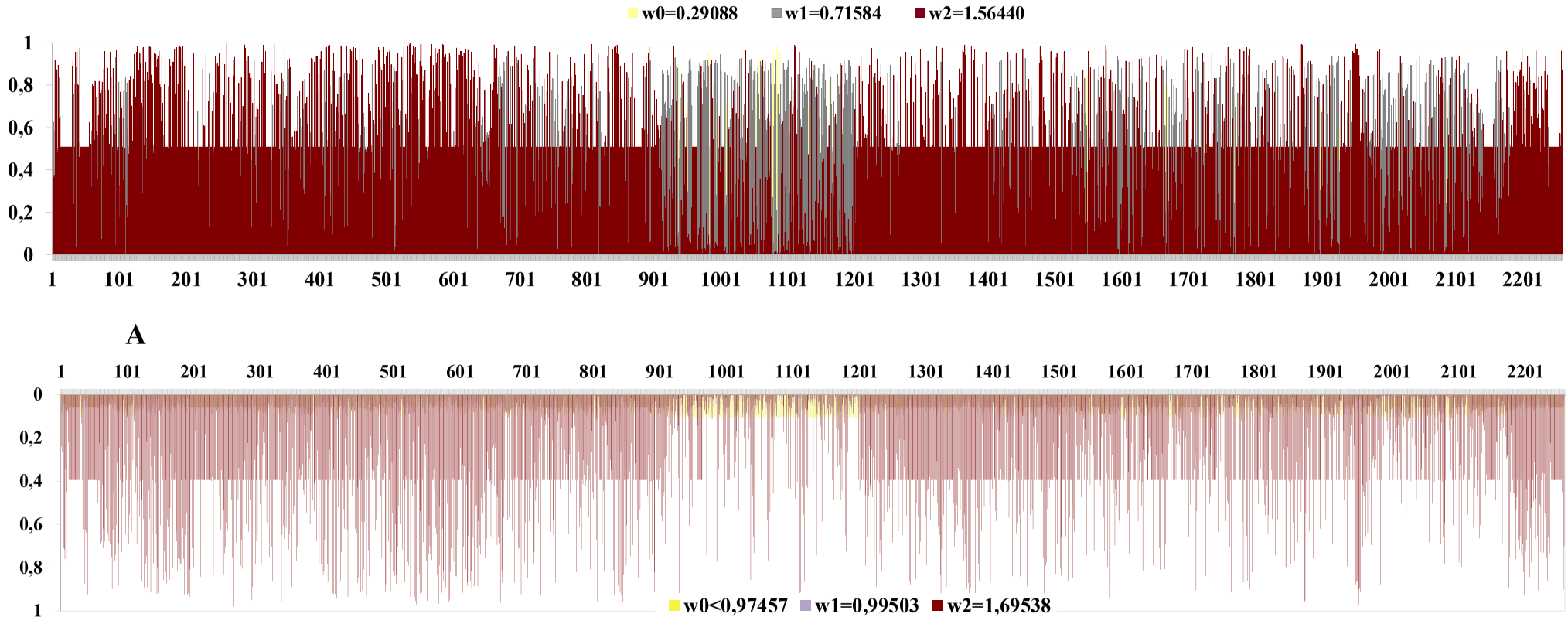


Figura 4.24-Probabilidade posterior das classes dos sítios da proteína gag-pol

(A) O modelo M3 utilizado assume três classes de sítios no gene. As frequências estimadas e as razões ω para as três classes são p_0 = 0,046, p_1 = 0,444 e p_2 = 0,508 e ω_0 = 0,290, ω_1 = 0,715, ω_2 = 1,564 (tabela 4.12). Essas *pk* são as distribuições a priori para cada sítio do códon (aminoácido). Os dados (configurações dos códons nas diferentes sequências) nos sítios mudam essa distribuição a priori para uma distribuição posterior, que pode ser muito diferente da a priori. Por exemplo, as probabilidades posteriores para as três classes no sítio 1 são 0,9997, 0,0002 e 0,0000, e o sítio está sob seleção purificadora forte. Em contraste, as probabilidades posteriores no sítio 262 são 0,0000, 0,0030 e 0,9970, e o sítio está certamente sob seleção de diversificação.

(B) O modelo M8 (beta&ω) se ajusta aos dados com onze categorias. As razões ω para as onze categorias são 0.33422, 0.52681, 0.64512, 0.73307, 0.80288, 0.85985, 0.90670, 0.94480, 0.97457, 0.99503 e 1.69538 (tabela 4.12). As primeiras nove categorias foram combinadas no gráfico.

4.3.2.7 Integrase

O modelo M2 sugeriu que aproximadamente 23% dos sítios estão sob seleção positiva com ω_2 = 4,388 (tabela 4.15). A estatística do teste LRT foi 2ΔlnL = 2 x ((-71156,59) - (-71173,81)) = 34,44 com *P* = 0,0008 e grau de liberdade (gl) = 2. Portanto, o modelo **M2** foi significantemente melhor que o modelo **M1** para os dados.

O modelo M3 sugeriu que 35% dos sítios estão sob seleção positiva com ω_2 = 2,805 e identificou 34 aminoácidos sob seleção positiva. A estatística do teste LRT foi 2ΔlnL = 2 x ((-71153,92) - (-56603,87)) = 29100,1 com *P* < 0,0001 e gl=4. O modelo **M3** foi extrema e significativamente melhor que o modelo de razão única **M0**.

O modelo M8 sugeriu que aproximadamente 30% dos sítios estão sob seleção positiva com ω = 3,161 e identificou 16 sítios sob seleção positiva (sítios sugeridos também por M3, tabela 4.15). A estatística do teste LRT foi 2ΔlnL = 53,16 com *P* = 0,0004 e gl=2. O modelo **M8** foi significantemente melhor que o modelo **M7**.

De acordo com Yang e Nielsen (2002) a comparação M0-M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção positiva. Sendo assim, a pressão seletiva sob a integrase parece ser muito variável ao longo de sua sequência. A distribuição de probabilidade posterior para os sítios da integrase com os resultados do modelo M3 é apresentada na figura 4.24.

Tabela 4.15 - Resultados do Teste de máxima verossimilhança para a seleção positiva na integrase

Modelo[1]	p[2]	*l*[3]	*k*[4]	Parâmetros estimados[5]		LRT[6]	Sítios positivamente selecionados[7]
M0	39	-56603,87	1,77	ω = 0,632		N/A	Nenhum
M1	40	-71173,81	1,91	p_0 = 0,360 p_1 = 0,639		N/A	Não permitido
M2	42	-71156,59	2,12	p_0 = 0,282 p_1 = 0,484 **p_2 = 0,232**	**ω_2 = 4,388**	&34,44 (*P* = 0,0008)	5H** 12D 15L 218A 929S 988S 1134Y
M3	43	-71153,92	2,13	p_0 = 0,151 p_1 = 0,498 **p_2 = 0,349**	ω_0 = 0,167 ω_1 = 0,598 **ω_2 = 2,805**	&&**29100,1** **(***P* < 0,0001**)**	4E 5H** 6V 12D 15L 65S 67I 69R 72V 74L 218A** 325Y 331N** 336N 338I 553M 737T 741S 742R 890G** 904P 929S** 934L** 937N 942S 947H 973Q** 988S** 1021H 1125K** 1133K 1134Y** 1142S 1258R
M7	40	-71179,25	1,91	p = 1,050 q = 0,692		N/A	Não permitido
M8	42	-71152,67	2,12	p_0 = 0,703 p = 1,491 q = 0,679	**p_1 = 0,296** **ω = 3,161**	&53,16 (*P* = 0,0004)	**4E 5H** 6V 12D 15L 218A** 331N 741S 890G 929S 934L 942S 988S 1125K 1134Y 1142S**

& Valor estatisticamente significativo para a o teste LRT (P ≤ 0.05)
&& Valor extrema e estatisticamente significativo para a o teste LRT (P ≤

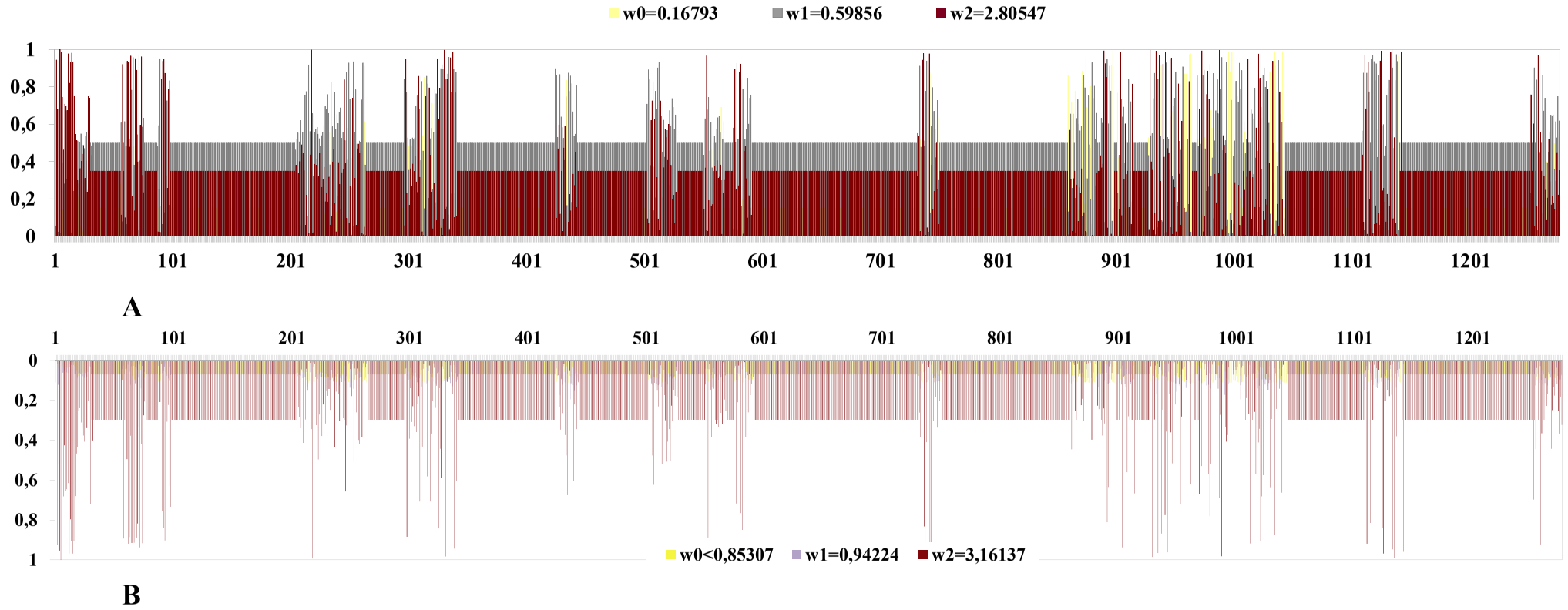


Figura 4.25-Probabilidade posterior das classes dos sítios da integrase

(A) O modelo M3 utilizado assume três classes de sítios no gene. As frequências estimadas e as razões ω para as três classes são p_0 = 0,151, p_1 = 0,4998 e p_2 = 0,349 e ω_0 = 0,167, ω_1 = 0,598, ω_2 = 2,805 (tabela 4.13). Essas *pk* são as distribuições a priori para cada sítio do códon (aminoácido). Os dados (configurações dos códons nas diferentes sequências) nos sítios mudam essa distribuição a priori para uma distribuição posterior, que pode ser muito diferente da a priori. Por exemplo, as probabilidades posteriores para as três classes no sítio 1 são 0,9978, 0,0022 e 0,0000, e o sítio está sob seleção purificadora forte. Em contraste, as probabilidades posteriores no sítio 5 são 0,0000, 0,001 e 0,9999, e o sítio está certamente sob seleção de diversificação.
(B) O modelo M8 (beta&ω) se ajusta aos dados com onze categorias. As razões ω para as onze categorias são 0.11549, 0.24412, 0.34751, 0.44004, 0.52646, 0.60925, 0.69013, 0.77069, 0.85307, 0.94224 e 3.16137 (tabela 4.13). As primeiras nove categorias foram combinadas no gráfico.

4.3.2.8 Ribonuclease H

O modelo M2 sugeriu que aproximadamente 2% dos sítios estão sob seleção positiva com ω_2 = 2,295 (tabela 4.16). A estatística do teste LRT foi 2ΔlnL = 2 x ((-105162,38) - (-105186,12)) = 47,48 com *P* = 0,0004 e grau de liberdade (gl) = 2. Portanto, o modelo **M2** foi significantemente melhor que o modelo **M1** para os dados.

O modelo M3 sugeriu que 61% dos sítios estão sob seleção positiva com ω_2 = 0,980, mas não identificou quais os sítios. A estatística do teste LRT foi 2ΔlnL = 2 x ((-105140,62) - (-69732,07)) = 70817,1 com *P* < 0,0001 e gl=4. O modelo **M3** foi extrema e significativamente melhor que o modelo de razão única **M0**.

O modelo sugeriu que aproximadamente 49% dos sítios estão sob seleção positiva com ω = 1,000 mas no identificou quais os sítios. No entanto, as diferenças entre M7 e M8 não são estatisticamente significativas. A estatística do teste LRT foi 2ΔlnL = 4,06 com *P* = 0,0557 e gl=2.

De acordo com Yang e Nielsen (2002) a comparação M0-M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção positiva. Sendo assim, a pressão seletiva sob a ribonuclease H parece ser muito variável ao longo de sua sequência. A distribuição de probabilidade posterior para os sítios da ribonuclease H com os resultados do modelo M3 é apresentada na figura 4.25.

Tabela 4.16 - Resultados do Teste de máxima verossimilhança para a seleção positiva na ribonuclease H

Modelo[1]	p[2]	*l*[3]	*k*[4]	Parâmetros estimados[5]		LRT[6]	Sítios positivamente selecionados[7]
M0	73	-69732,07	1,99	ω = 0,725		N/A	Nenhum
M1	74	-105186,12	2,06	p_0 = 0,002 p_1 = 0,997		N/A	Não permitido
M2	76	-105162,38	2,15	p_0 = 0,330 p_1 = 0,650 **p_2 = 0,018**	**ω_2 = 2,295**	[&]47,48, (*P* = 0,0004)	Nenhum
M3	77	-105140,62	2,14	p_0 = 0,002 p_1 = 0,391 **p_2 = 0,606**	ω_0 = 0,000 ω_1 = 0,493 **ω_2 = 0,980**	[&&]**70817,1** **(***P* < 0,0001**)**	Nenhum
M7	74	-105159,36	2,10	p = 2,235 q = 0,745		N/A	Não permitido
M8	76	-105157,33	2,13	p_0 = 0,507 p = 4,134 q = 3,014	**p_1 = 0,492** **ω = 1,000**	4,06 (*P* = 0,0557)	**Nenhum**

[&] Valor estatisticamente significativo para a o teste LRT (P ≤ 0.05)
[&&] Valor extrema e estatisticamente significativo para a o teste LRT (P ≤ 0.0001)

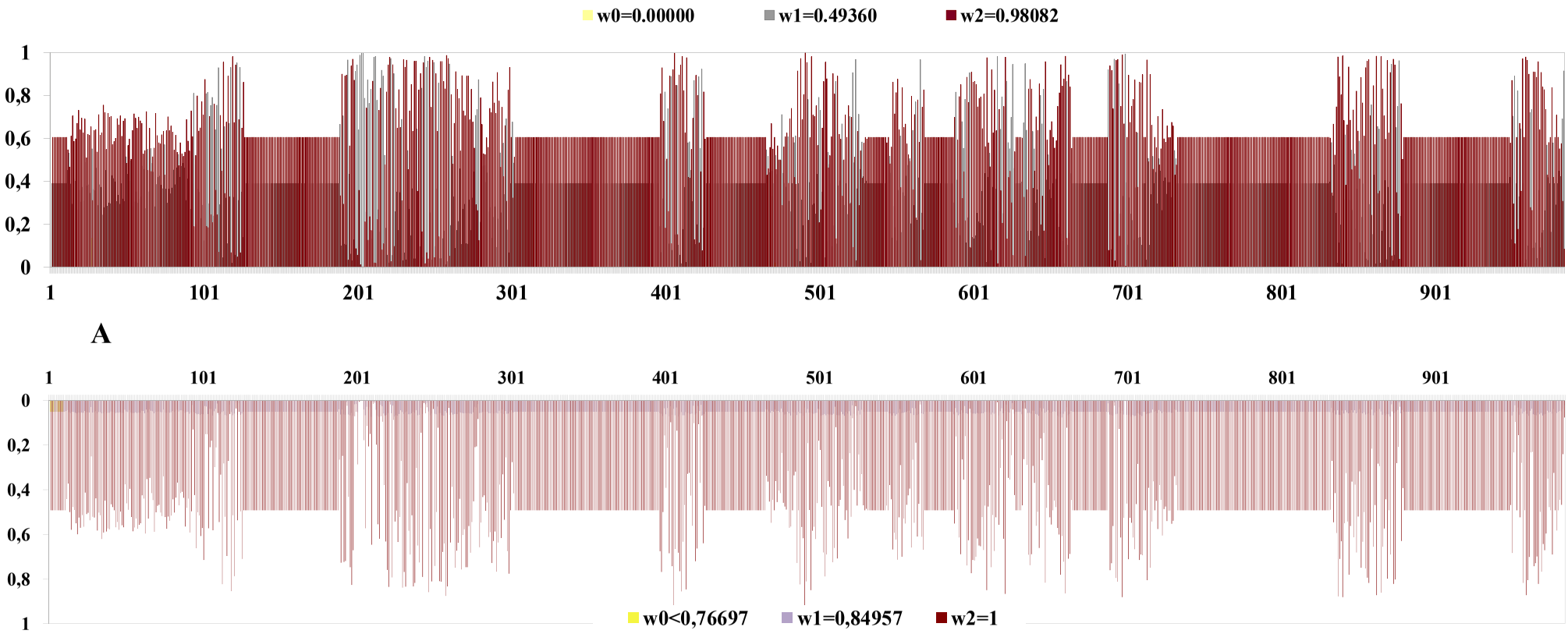


Figura 4.26-Probabilidade posterior das classes dos sítios da ribonuclease H

(A) O modelo M3 utilizado assume três classes de sítios no gene. As frequências estimadas e as razões ω para as três classes são $p_0 = 0{,}002$, $p_1 = 0{,}391$ e $p_2 = 0{,}606$ e $\omega_0 = 0{,}000$, $\omega_1 = 0{,}493$, $\omega_2 = 0{,}980$ (tabela 4.14). Essas *pk* são as distribuições a priori para cada sítio do códon (aminoácido). Os dados (configurações dos códons nas diferentes sequências) nos sítios mudam essa distribuição a priori para uma distribuição posterior, que pode ser muito diferente da a priori. Por exemplo, as probabilidades posteriores para as três classes no sítio 491 são 0,0000, 0,019 e 0,9981, e o sítio está certamente sob seleção de diversificação.
(B) O modelo M8 (beta&ω) se ajusta aos dados com onze categorias. As razões ω para as onze categorias são 0.28059, 0.38711, 0.45565, 0.51165, 0.56191, 0.60980, 0.65775, 0.70852, 0.76697, 0.84957 e 1.00000 (tabela 4.14). As primeiras nove categorias foram combinadas no gráfico.

Legenda das tabelas 4.9-4.16

[1] Modelos de sequência de evolução de códons baseados em máxima verossimilhança implementados no PAML: M0, razão única; M1, neutro; M2, seleção; M3, discreto; M7, beta; M8, beta + ω .

[2] p = número de parâmetros nos cálculos de máxima verossimilhança.

[3] *l* = *log likelihood* calculados pelo codeml.

[4] κ = Kappa (taxa da razão de transição/transversão), estimado dos dados pelo codeml .

[5] $\omega = K_A/K_S$ razão estimada dos dados (M0, um ω estimado; M1, ω = 0 or 1; M2, ω = 0, 1 ou a terceira classe do sítio estimado; M3, ω_1, ω_2, e ω_3 estimados dos dados; M7, ω = distribuição beta entre 0 e 1; M8, como para M7 exceto para a classe do sítio adicional onde um ω estimado é permitido); p = proporção de sítios por classe; *p*,*q* = parâmetros da distribuição beta.

[6] LRT = teste do radio de verossimilhança que compara o teste estatístico ($2\Delta l$) calculado dos modelos pareados (exemplo: M1/M2, M0/M3 e M7/M8) com o valor crítico de χ^2 com os graus de liberdades correspondentes (exemplo: 2 g.l., 4 g.l. e 2 g.l., respectivamente). $2\Delta l$ e a significância são mostrados para M2, M3 e M8. Abreviaturas: N/A, não aplicável.

[7] Sítios selecionados positivamente com probabilidade posterior (*P*) > 0.5 são mostrados. Resíduos com uma seleção inferida no nível de 99% são marcados **, Resíduos com uma seleção inferida no nível de 95% são marcados *. Os números dos resíduos correspondem às colunas das sequências em referência.

4.4 A telomerase e os elementos de retrotransposição em Tri-tryps baseados em análises filogenéticas dos domínios da transcriptase reversa

A adição dos alinhamentos dos grupos **ia**, **ib** e **ii** ao alinhamento PF00078-RVT_1 resultou em um alinhamento formado por 227 táxons de espécies diferentes mostrando os oito domínios de 440 aminoácidos que foi reduzido a 88 táxons.

4.4.1 Análises filogenéticas

A figura 4.28 mostra o alinhamento múltiplo de sequências de aminoácidos dos oito domínios que caracterizam a transcriptase reversa. Globalmente, este alinhamento, é muito similar ao mostrado por Xiong e Eickbush em 1990. A maioria das transcriptases reversas mostrou uma considerável divergência nas sequências.

A figura 4.27 mostra o fluxograma do algoritmo utilizado na execução dos programas de filogenia. As diferenças encontradas entre as árvores filogenéticas da transcriptase reversa geradas pelos programas PAUP-AV, PAUP-MP, PHYML, MRBAYES e WEIGHBOR basearam-se principalmente nos valores de *bootstrap* e as diferentes topologias das árvores encontradas. Por conseguinte, as árvores com maior consistência foram às construídas com os programas PAUP-AV, PAUP-MP e WEIGHBOR.

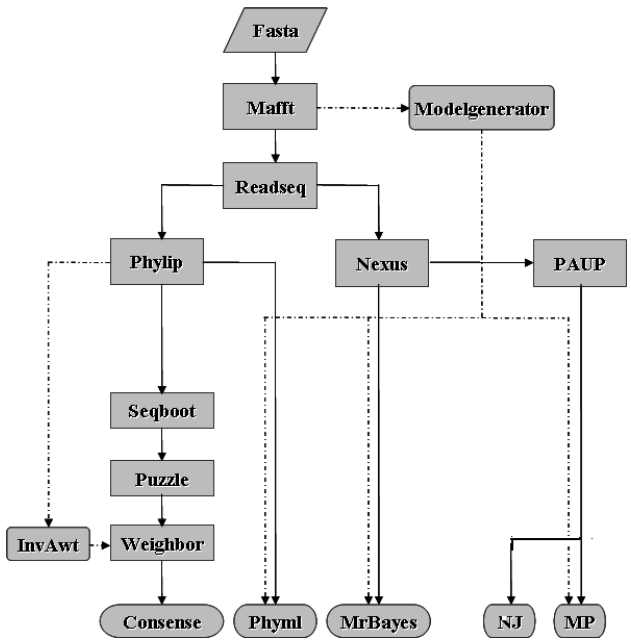
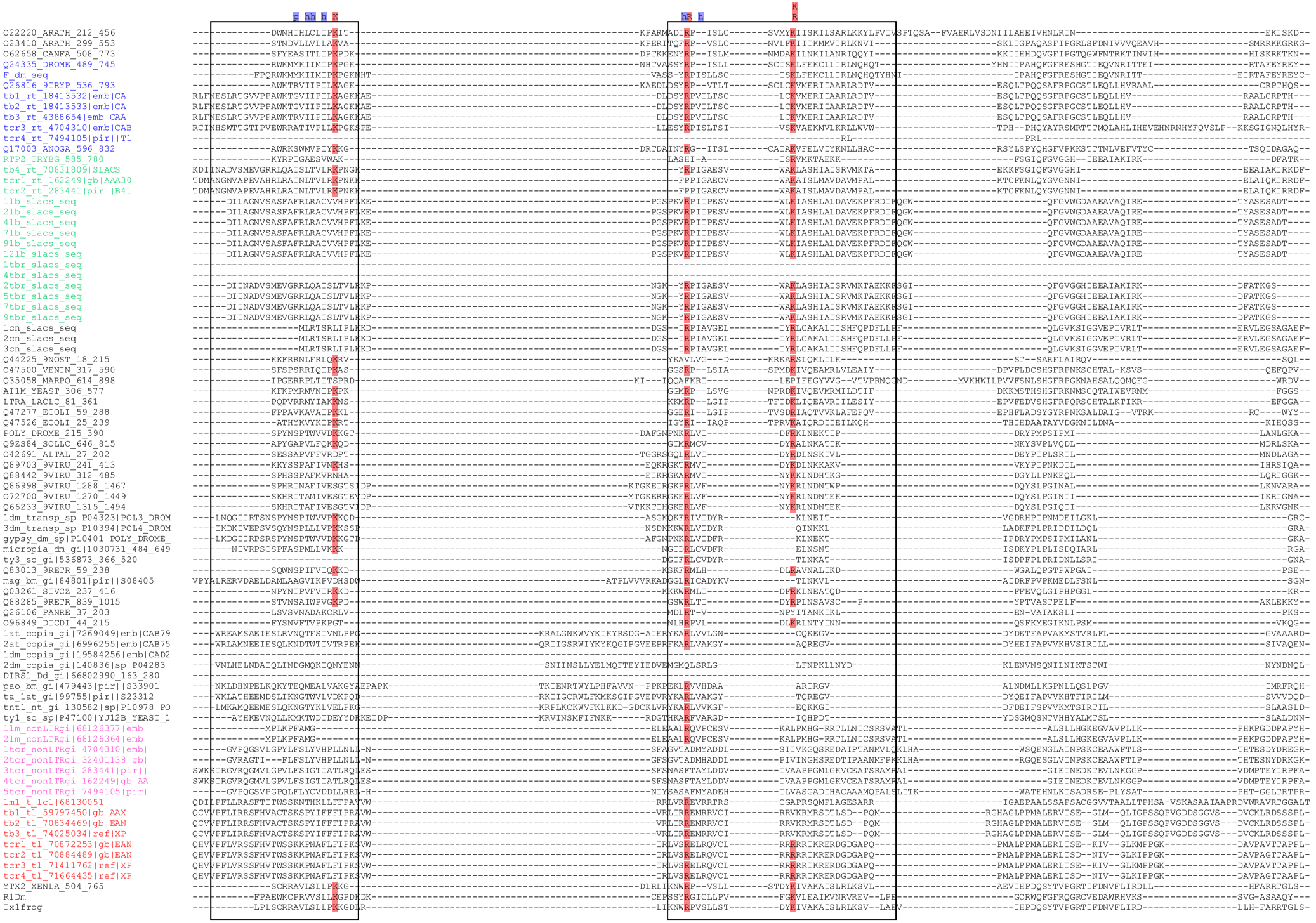
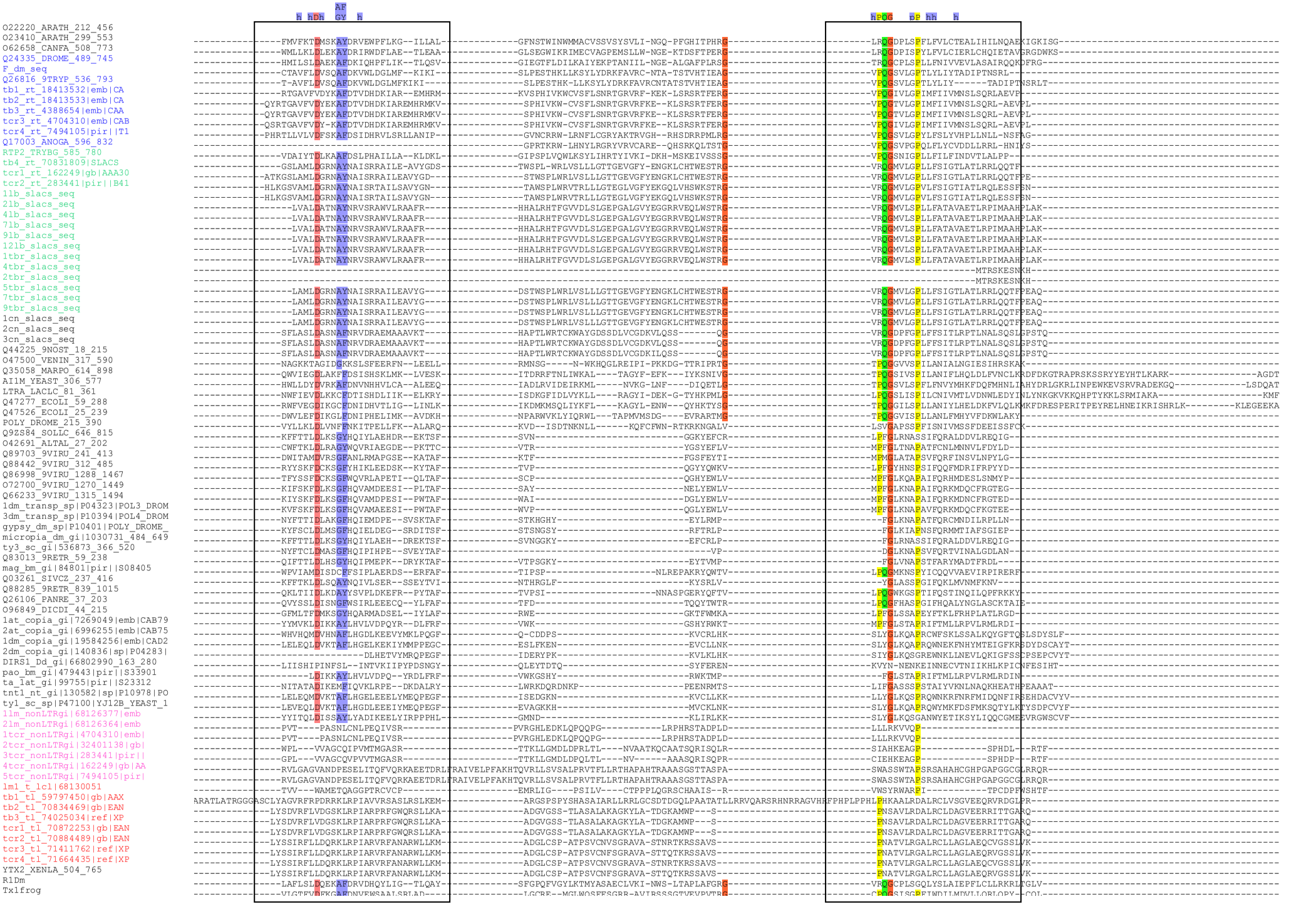
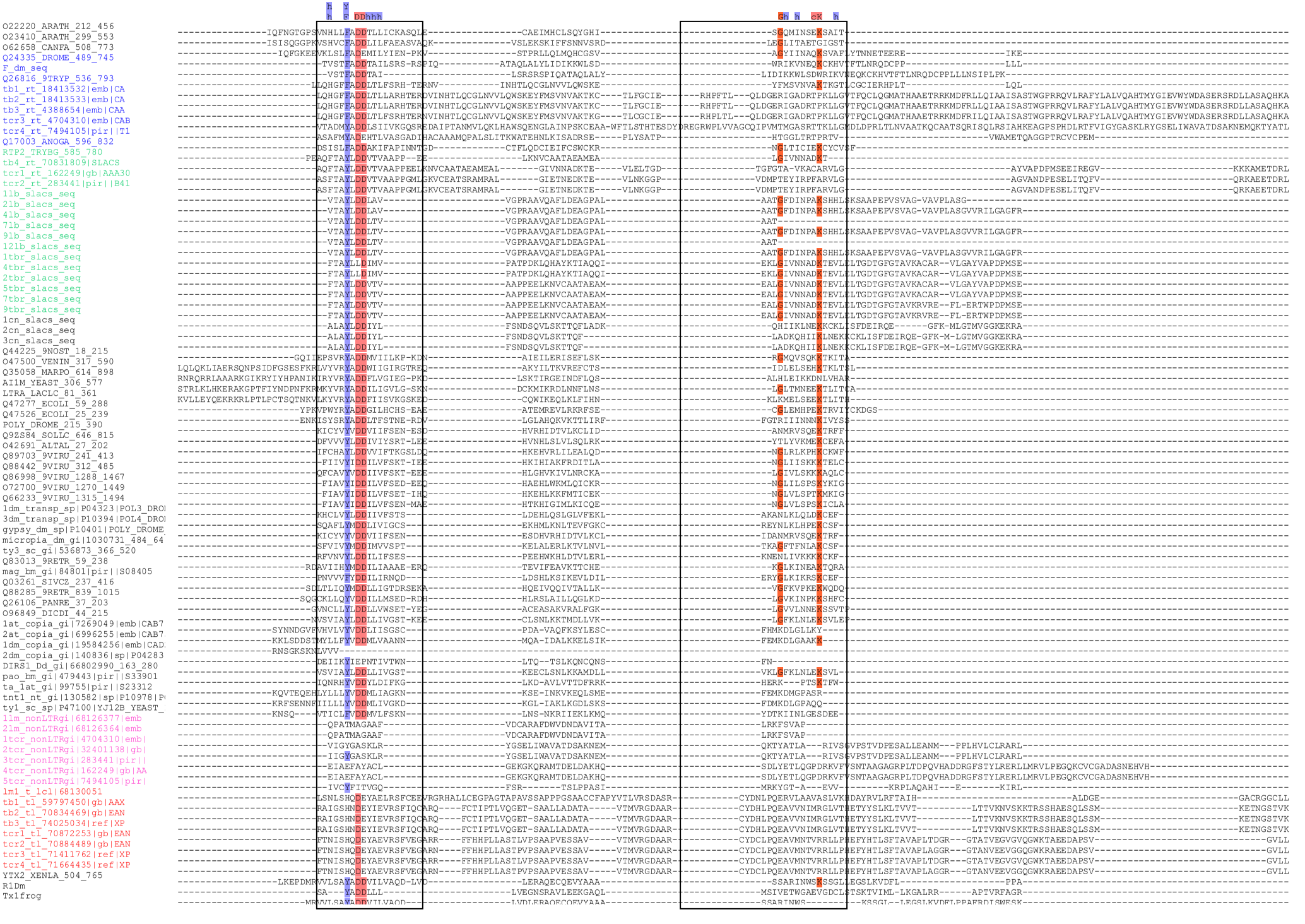


Figura 4.27 - Fluxograma usado para as execuções automáticas dos algoritmos filogenéticos


O22220_ARATH_212_456
O23410_ARATH_299_553
O62658_CANFA_508_773
Q24335_DROME_489_745
F_dm_seq
Q26816_9TRYP_536_793
tb1_rt_18413532|emb|CA
tb2_rt_18413533|emb|CA
tb3_rt_4388654|emb|CAA
tcr3_rt_4704310|emb|CAB
tcr4_rt_7494105|pir||T1
Q17003_ANOGA_596_832
RTP2_TRYBG_585_780
tb4_rt_70831809|SLACS
tcr1_rt_162249|gb|AAA30
tcr2_rt_283441|pir||B41
1lb_slacs_seq
2lb_slacs_seq
4lb_slacs_seq
7lb_slacs_seq
9lb_slacs_seq
12lb_slacs_seq
1tbr_slacs_seq
4tbr_slacs_seq
2tbr_slacs_seq
5tbr_slacs_seq
7tbr_slacs_seq
9tbr_slacs_seq
1cn_slacs_seq
2cn_slacs_seq
3cn_slacs_seq
Q44225_9NOST_18_215
O47500_VENIN_317_590
Q35058_MARPO_614_898
AI1M_YEAST_306_577
LTRA_LACLC_81_361
Q47277_ECOLI_59_288
Q47526_ECOLI_25_239
POLY_DROME_215_390
Q9ZS84_SOLLC_646_815
O42691_ALTAL_27_202
Q89703_9VIRU_241_413
Q88442_9VIRU_312_485
Q86998_9VIRU_1288_1467
O72700_9VIRU_1270_1449
Q66233_9VIRU_1315_1494
1dm_transp_sp|P04323|POL3_DROM
3dm_transp_sp|P10394|POL4_DROM
gypsy_dm_sp|P10401|POLY_DROME_
micropia_dm_gi|1030731_484_649
ty3_sc_gi|536873_366_520
Q83013_9RETR_59_238
mag_bm_gi|84801|pir||S08405
Q03261_SIVCZ_237_416
Q88285_9RETR_839_1015
Q26106_PANRE_37_203
O96849_DICDI_44_215
1at_copia_gi|7269049|emb|CAB79
2at_copia_gi|6996255|emb|CAB75
1dm_copia_gi|19584256|emb|CAD2
2dm_copia_gi|140836|sp|P04283|
DIRS1_Dd_gi|66802990_163_280
pao_bm_gi|479443|pir||S33901
ta_1at_gi|99755|pir||S23312
tnt1_nt_gi|130582|sp|P10978|PO
ty1_sc_sp|P47100|YJ12B_YEAST_1
1lm_nonLTRgi|68126377|emb
2lm_nonLTRgi|68126364|emb
1tcr_nonLTRgi|4704310|emb|
2tcr_nonLTRgi|32401138|gb|
3tcr_nonLTRgi|283441|pir||
4tcr_nonLTRgi|162249|gb|AA
5tcr_nonLTRgi|7494105|pir|
lm1_t_lcl|68130051
tb1_tl_59797450|gb|AAX
tb2_tl_70834469|gb|EAN
tb3_tl_74025034|ref|XP
tcr1_tl_70872253|gb|EAN
tcr2_tl_70884489|gb|EAN
tcr3_tl_71411762|ref|XP
tcr4_tl_71664435|ref|XP
YTX2_XENLA_504_765
R1Dm
Tx1frog


O22220_ARATH_212_456
O23410_ARATH_299_553
O62658_CANFA_508_773
Q24335_DROME_489_745
F_dm_seq
Q26816_9TRYP_536_793
tb1_rt_18413532|emb|CA
tb2_rt_18413533|emb|CA
tb3_rt_4388654|emb|CAA
tcr3_rt_4704310|emb|CAB
tcr4_rt_7494105|pir||T1
Q17003_ANOGA_596_832
RTP2_TRYBG_585_780
tb4_rt_70831809|SLACS
tcr1_rt_162249|gb|AAA30
tcr2_rt_283441|pir||B41
11b_slacs_seq
21b_slacs_seq
41b_slacs_seq
71b_slacs_seq
91b_slacs_seq
121b_slacs_seq
1tbr_slacs_seq
4tbr_slacs_seq
2tbr_slacs_seq
5tbr_slacs_seq
7tbr_slacs_seq
9tbr_slacs_seq
1cn_slacs_seq
2cn_slacs_seq
3cn_slacs_seq
Q44225_9NOST_18_215
O47500_VENIN_317_590
Q35058_MARPO_614_898
AI1M_YEAST_306_577
LTRA_LACLC_81_361
Q47277_ECOLI_59_288
Q47526_ECOLI_25_239
POLY_DROME_215_390
Q9ZS84_SOLLC_646_815
O42691_ALTAL_27_202
Q89703_9VIRU_241_413
Q88442_9VIRU_312_485
Q86998_9VIRU_1288_1467
O72700_9VIRU_1270_1449
Q66233_9VIRU_1315_1494
1dm_transp_sp|P04323|POL3_DROM
3dm_transp_sp|P10394|POL4_DROM
gypsy_dm_sp|P10401|POLY_DROME_
micropia_dm_gi|1030731_484_649
ty3_sc_gi|536873_366_520
Q83013_9RETR_59_238
mag_bm_gi|84801|pir||S08405
Q03261_SIVCZ_237_416
Q88285_9RETR_839_1015
Q26106_PANRE_37_203
O96849_DICDI_44_215
1at_copia_gi|7269049|emb|CAB79
2at_copia_gi|6996255|emb|CAB75
1dm_copia_gi|19584256|emb|CAD2
2dm_copia_gi|140836|sp|P04283|
DIRS1_Dd_gi|66802990_163_280
pao_bm_gi|479443|pir||S33901
ta_1at_gi|99755|pir||S23312
tnt1_nt_gi|130582|sp|P10978|PO
ty1_sc_sp|P47100|YJ12B_YEAST_1
11m_nonLTRgi|68126377|emb
21m_nonLTRgi|68126364|emb
1tcr_nonLTRgi|4704310|emb|
2tcr_nonLTRgi|32401138|gb|
3tcr_nonLTRgi|283441|pir||
4tcr_nonLTRgi|162249|gb|AA
5tcr_nonLTRgi|7494105|pir|
1m1_t_1cl|68130051
tb1_tl_59797450|gb|AAX
tb2_tl_70834469|gb|EAN
tb3_tl_74025034|ref|XP
tcr1_tl_70872253|gb|EAN
tcr2_tl_70884489|gb|EAN
tcr3_tl_71411762|ref|XP
tcr4_tl_71664435|ref|XP
YTX2_XENLA_504_765
R1Dm
Tx1frog


O22220_ARATH_212_456
O23410_ARATH_299_553
O62658_CANFA_508_773
Q24335_DROME_489_745
F_dm_seq
Q26816_9TRYP_536_793
tb1_rt_18413532|emb|CA
tb2_rt_18413533|emb|CA
tb3_rt_4388654|emb|CAA
tcr3_rt_4704310|emb|CAB
tcr4_rt_7494105|pir||T1
Q17003_ANOGA_596_832
RTP2_TRYBG_585_780
tb4_rt_70831809|SLACS
tcr1_rt_162249|gb|AAA30
tcr2_rt_283441|pir||B41
11b_slacs_seq
21b_slacs_seq
41b_slacs_seq
71b_slacs_seq
91b_slacs_seq
121b_slacs_seq
1tbr_slacs_seq
4tbr_slacs_seq
2tbr_slacs_seq
5tbr_slacs_seq
7tbr_slacs_seq
9tbr_slacs_seq
1cn_slacs_seq
2cn_slacs_seq
3cn_slacs_seq
Q44225_9NOST_18_215
O47500_VENIN_317_590
Q35058_MARPO_614_898
AI1M_YEAST_306_577
LTRA_LACLC_81_361
Q47277_ECOLI_59_288
Q47526_ECOLI_25_239
POLY_DROME_215_390
Q9ZS84_SOLLC_646_815
O42691_ALTAL_27_202
Q89703_9VIRU_241_413
Q88442_9VIRU_312_485
Q86998_9VIRU_1288_1467
O72700_9VIRU_1270_1449
Q66233_9VIRU_1315_1494
1dm_transp_sp|P04323|POL3_DROI
3dm_transp_sp|P10394|POL4_DROI
gypsy_dm_sp|P10401|POLY_DROME_
micropia_dm_gi|1030731_484_64
ty3_sc_gi|536873_366_520
Q83013_9RETR_59_238
mag_bm_gi|84801|pir||S08405
Q03261_SIVCZ_237_416
Q88285_9RETR_839_1015
Q26106_PANRE_37_203
O96849_DICDI_44_215
1at_copia_gi|7269049|emb|CAB7
2at_copia_gi|6996255|emb|CAB7.
1dm_copia_gi|19584256|emb|CAD.
2dm_copia_gi|140836|sp|P04283
DIRS1_Dd_gi|66802990_163_280
pao_bm_gi|479443|pir||S33901
ta_1at_gi|99755|pir||S23312
tnt1_nt_gi|130582|sp|P10978|P
ty1_sc_gi|P47100|YJ12B_YEAST_
11m_nonLTRgi|68126377|emb
21m_nonLTRgi|68126364|emb
1tcr_nonLTRgi|4704310|emb|
2tcr_nonLTRgi|32401138|gb|
3tcr_nonLTRgi|283441|pir||
4tcr_nonLTRgi|162249|gb|AA
5tcr_nonLTRgi|7494105|pir|
1m1_t_1cl|68130051
tb1_tl_59797450|gb|AAX
tb2_tl_70834469|gb|EAN
tb3_tl_74025034|ref|XP
tcr1_tl_70872253|gb|EAN
tcr2_tl_70884489|gb|EAN
tcr3_tl_71411762|ref|XP
tcr4_tl_71664435|ref|XP
YTX2_XENLA_504_765
R1Dm
Tx1frog

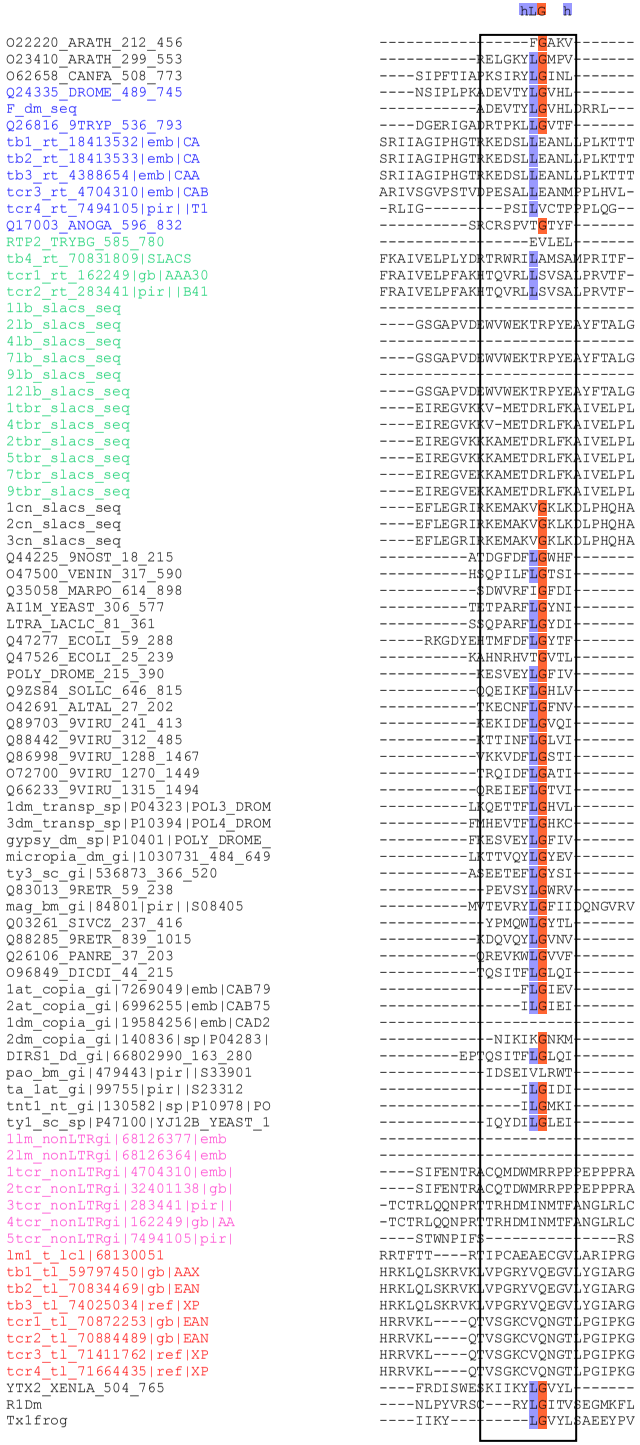


Figura 4.28 - Alinhamento baseado nos 440 aminoácidos pertencentes aos oito domínios da transcriptase reversa

Foram usadas 84 sequências das transcriptases reversas obtidas do banco de dados NCBI. O alinhamento foi obtido após execução do programa MAFFT.

Na figuras 4.29 e 4.31, observou-se que a topologia da árvore construída com o programa WEIGHBOR foi idêntica à árvore de construída com o PAUP-AV e a construída com o MRBAYES o foi com a construída com o PAUP-MP. Por outro lado, a árvore construída com o PHYML apresentou a topologia mais diferente quando comparada com as construídas com WEIGHBOR, PAUP-AV, PAUP-MP e MRBAYES. A árvore PAUP-AV foi escolhida para realizar as descrições da filogenia e as outras árvores foram utilizadas no esclarecimento de alguma dúvida ou incoerência, ou para reforçar a inferência.

Com base nas análises filogenéticas, as transcriptases reversas dos EGM retrotransponíveis em Tri-tryps foram agrupadas em quatro clados pertencentes a (figura 4.29 e 4.31):

- Clado 1: os retrotransposons não-LTR (cor azul),
- Clado 2: os retrotransposons não-LTR SLACS, as ORF, as sequências hipotéticas e a outras transcriptases reversas. Mais de um terço destas transcriptases reversas foi anotado de forma inespecífica, sem informação sobre o retrotransposon ao que pertencem (cor verde),
- Clado 3: os retrotransposons não-LTR (cor de rosa) e
- Clado 4: as telomerases. Este clado formou um ramo isolado externo monofilético (cor vermelha).

As análises filogenéticas indicaram monofilia para Tri-tryps em cada um dos clados, como pode ser observado nas árvores filogenéticas construídas com sequências de aminoácidos e nucleotídeos da transcriptase reversa (figura 4.30). Não foram identificadas transcriptases reversas pertencentes aos retrotransposons LTR em Tri-tryps.

O Clado 1 constituiu um grupo monofilético que contém sequências de retrotransposons não-LTR nas espécies *T. cruzi* (tcr3 e tcr4) que foram as mais externas e *T. brucei* (tbr1, tbr2 e tbr3).

O Clado 2 constituiu um grupo monofilético que contém sequências de retrotransposons SLACS nas espécies *L. brasiliensis*, *T. brucei*, *T. brucei gambiense* e *T. cruzi*. A árvore PAUP-AV mostrou dois subgrupos parafiléticos:

- subgrupo 1 formado por *L. brasiliensis* (1lb, 2lb, 4lb, 7lb, 9lb e 12lb) e
- subgrupo 2 formado por:

* *T. brucei* - SLACS, transcriptases reversas putativas - (1tbr, 2tbr, 4tbr, 5tbr, 7tbr, 9tbr),

* *T. brucei gambiense* - RTP2_TRYBG_585_780 SLACS - (tbg) e

* *T. cruzi* - endonuclease transcriptase reversa (tcr1), proteína 2 hipotética relacionada ao retrovírus-retrotransposon (tcr2)-.

O Clado 3 constituiu um grupo monofilético que contém sequências de retrotransposons não-LTR nas espécies *L. major* (1lm e 2lm) que foi a mais externa e *T. cruzi* (1tcr, 2tcr, 3tcr, 4tcr e 5tcr).

O Clado 4 constituiu um grupo monofilético que contém sequências de telomerases nas espécies *L. major* (lm1 tl) que foi a mais externa, *T. brucei* (tbr1 tl, tbr2 tl e tbr3 tl) e *T. cruzi* (tcr1 tl, tcr2 tl, tcr3 tl e tcr4 tl).

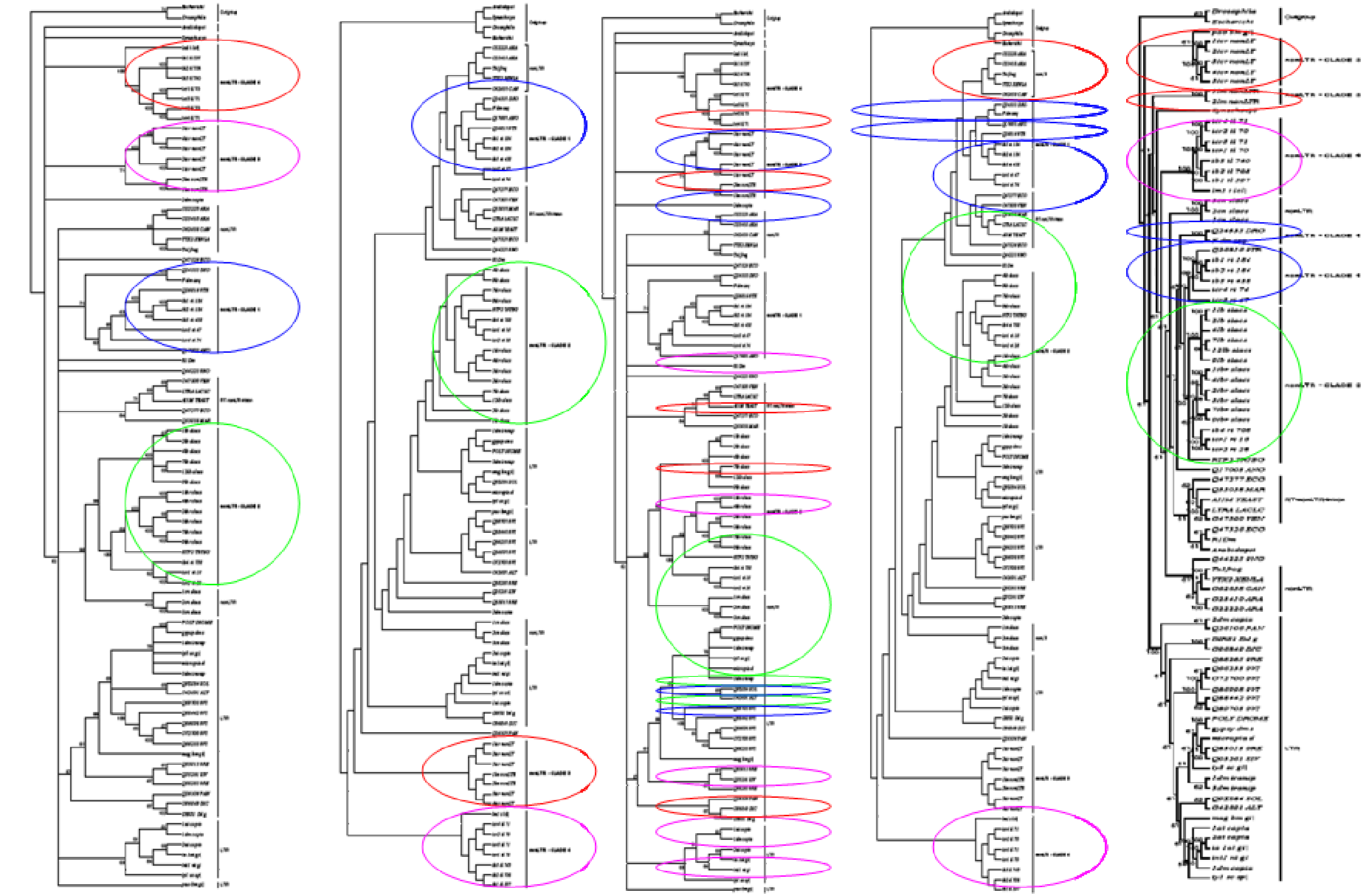


Figura 4.29 - Árvores filogenéticas da transcriptase reversa em Tri-tryps

As árvores filogenéticas foram construídas usando os algoritmos: (i) AV com o programa PAUP-AV, (ii) WEIGHBOR, (iii) MV com o PHYML, (iv) IB com o MRBAYES e (v) MP com o PAUP-MP.
Círculos de diferentes cores foram assinados para os quatro clados: (A) o azul para o Clado 1 dos não-LTR, (B) o verde para o Clado 2 dos não-LTR, (C) o cor de rosa para o Clado 3 dos não-LTR e (D) o vermelho para o Clado 4 das telomerases.

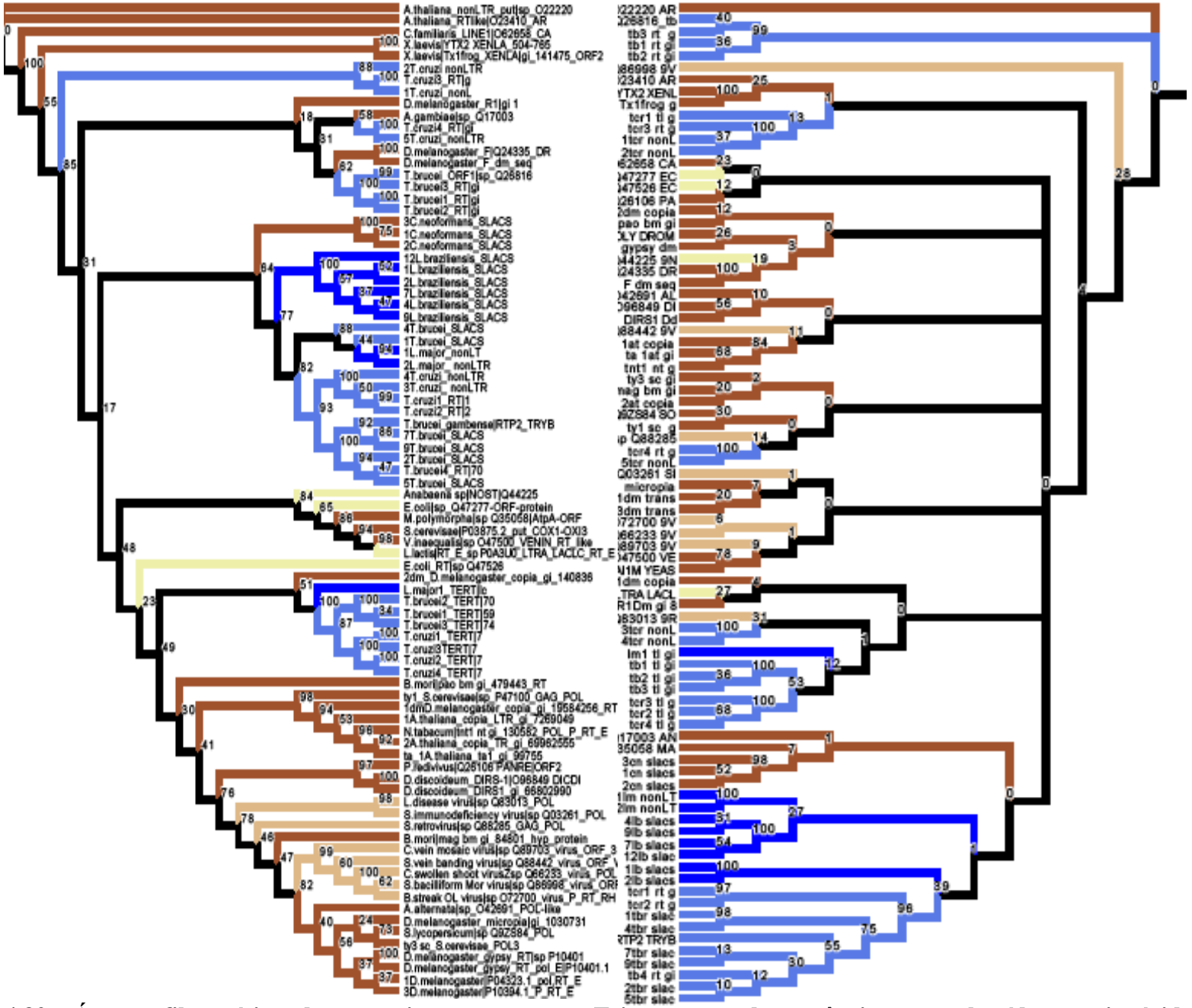
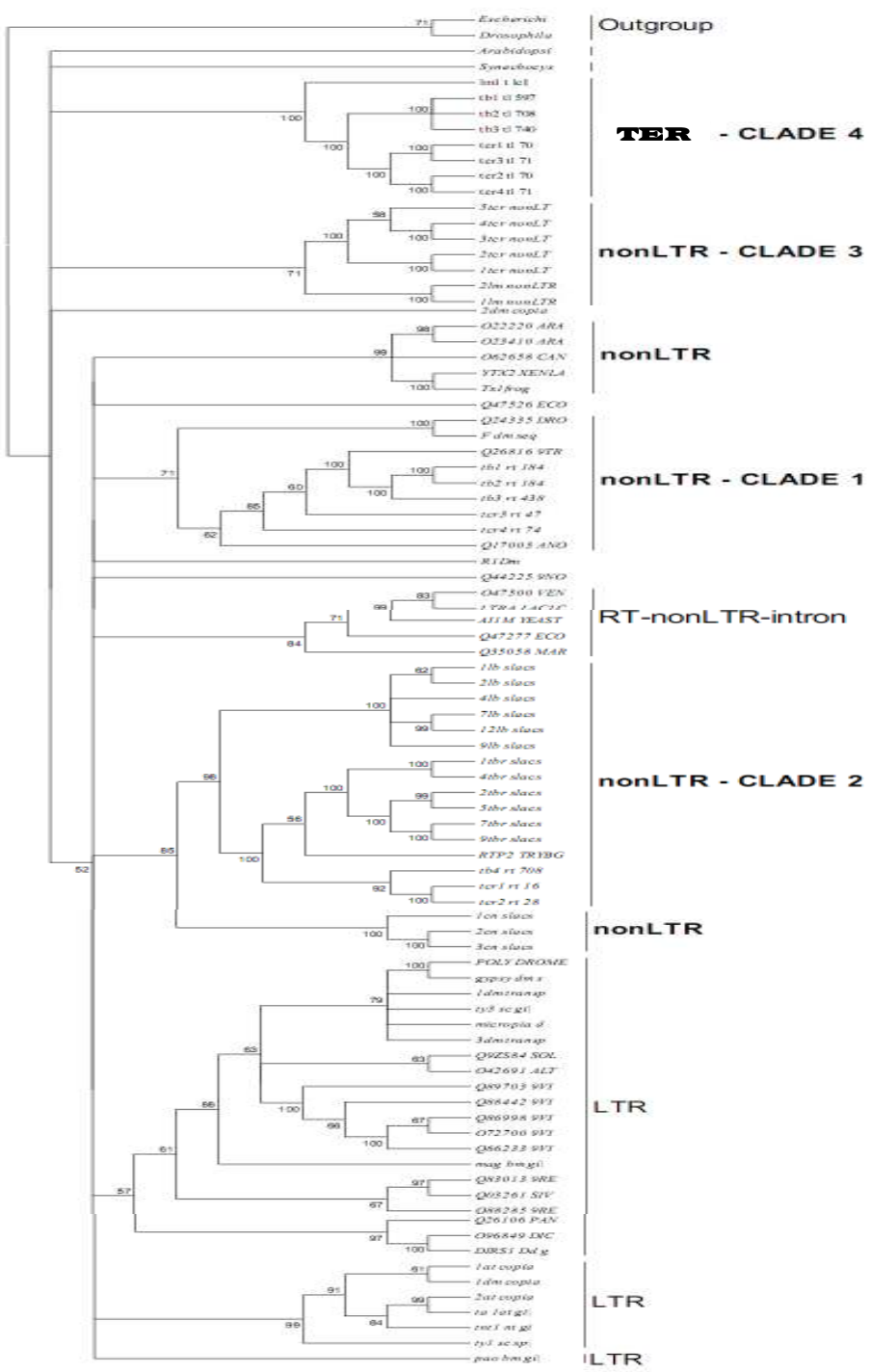
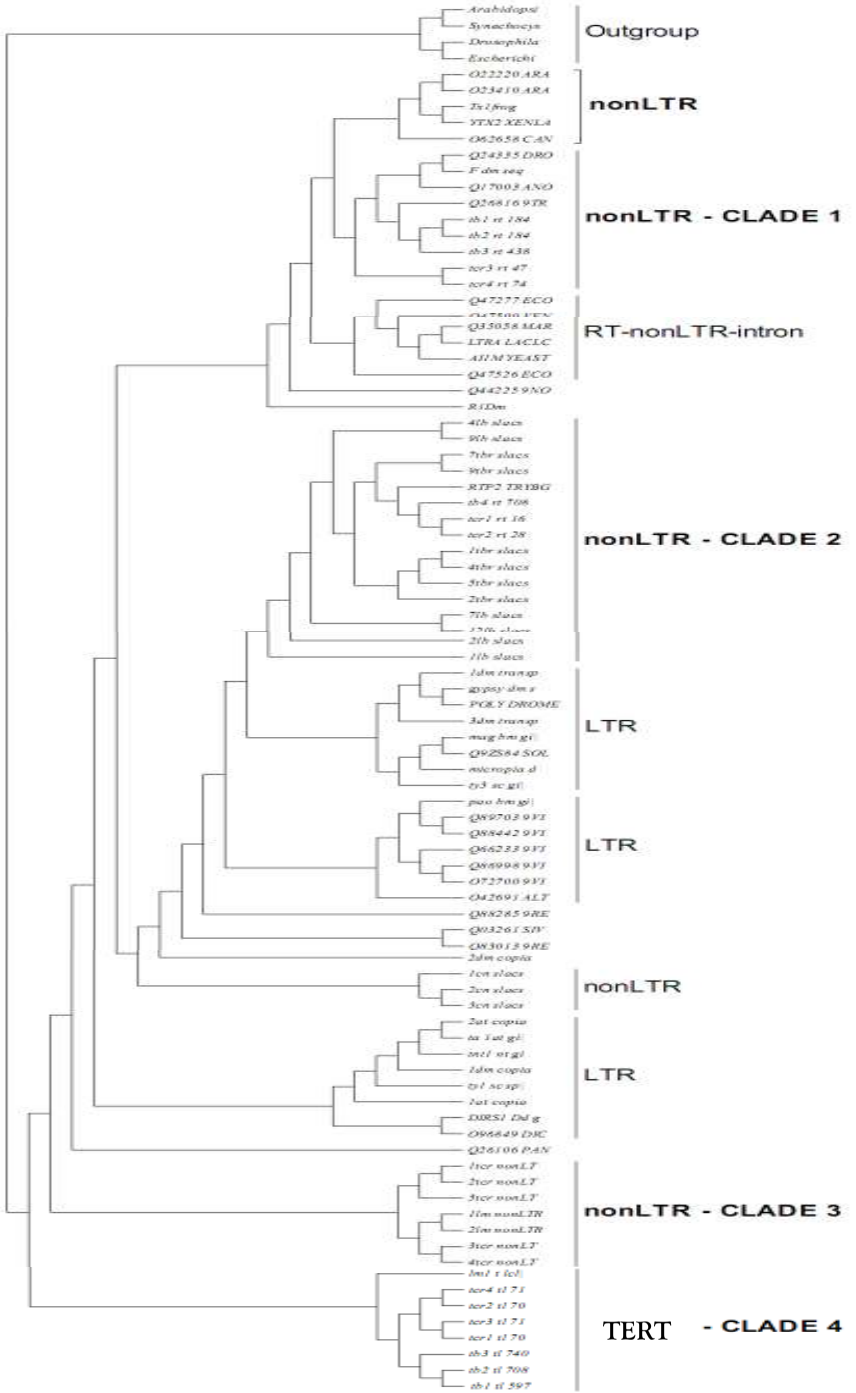


Figura 4.30 - Árvores filogenéticas da transcriptase reversa em Tri-tryps usando sequências em nucleotídeo e aminoácido

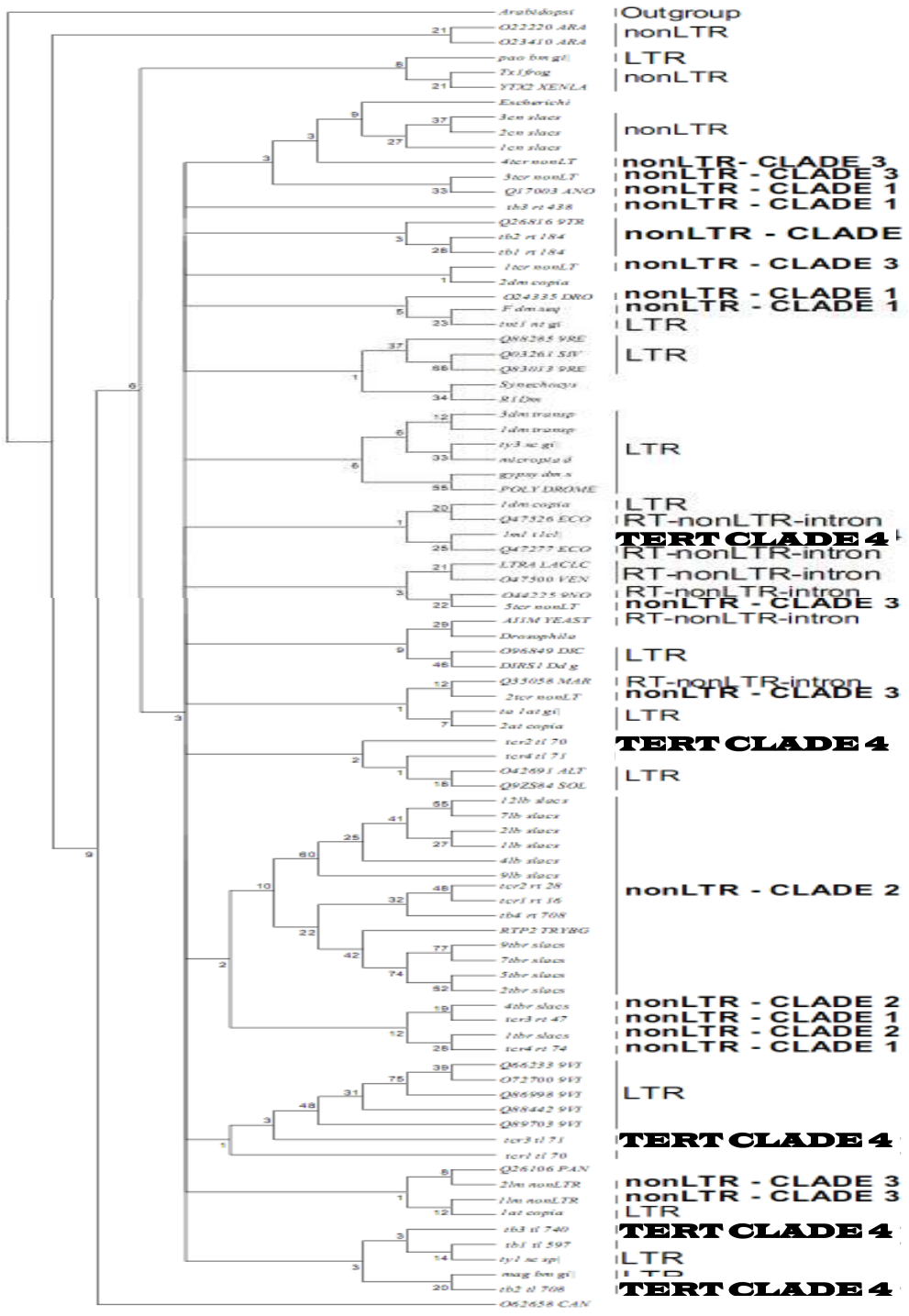
As árvores filogenéticas foram construídas usando o programa PAUP-AV, com um valor de *boostrap* de 10.000.


Outgroup
TER - CLADE 4
nonLTR - CLADE 3
nonLTR
nonLTR - CLADE 1
RT-nonLTR-intron
nonLTR - CLADE 2
nonLTR
LTR
LTR
LTR

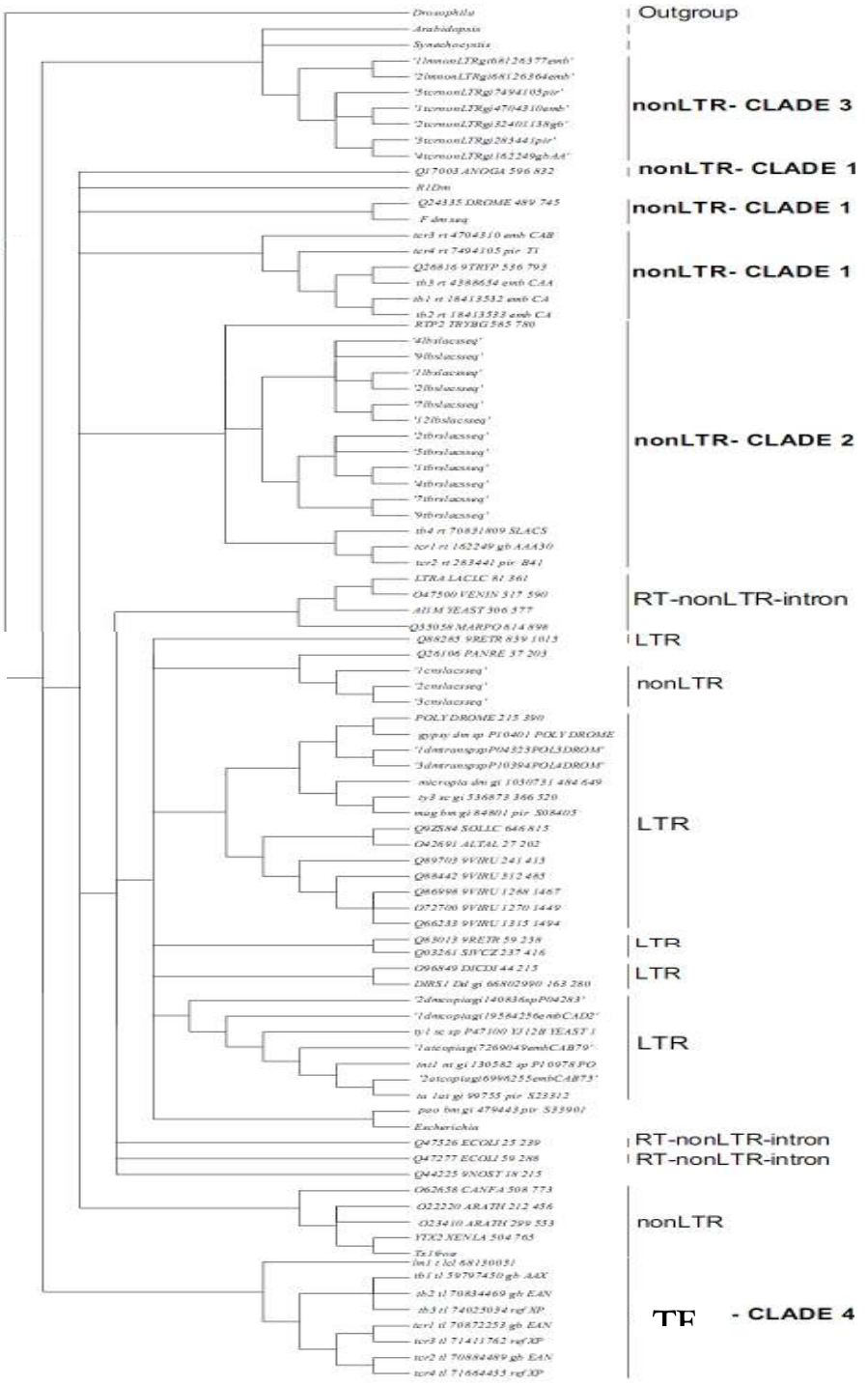
I


Outgroup
nonLTR
nonLTR - CLADE 1
RT-nonLTR-intron
nonLTR - CLADE 2
LTR
LTR
nonLTR
LTR
nonLTR - CLADE 3
TERT - CLADE 4

II.


Outgroup
nonLTR
LTR
nonLTR
nonLTR
nonLTR- CLADE 3
nonLTR - CLADE 3
nonLTR - CLADE 1
nonLTR - CLADE 1
nonLTR - CLADE
nonLTR - CLADE 3
nonLTR - CLADE 1
nonLTR - CLADE 1
LTR
LTR
LTR
RT-nonLTR-intron
TERT CLADE 4
RT-nonLTR-intron
RT-nonLTR-intron
RT-nonLTR-intron
nonLTR - CLADE 3
RT-nonLTR-intron
LTR
nonLTR - CLADE 3
LTR
TERT CLADE 4
LTR
nonLTR - CLADE 2
nonLTR - CLADE 2
nonLTR - CLADE 1
nonLTR - CLADE 2
nonLTR - CLADE 1
LTR
TERT CLADE 4
nonLTR - CLADE 3
nonLTR - CLADE 3
LTR
TERT CLADE 4
LTR
TERT CLADE 4

III.


Outgroup
nonLTR- CLADE 3
nonLTR- CLADE 1
nonLTR- CLADE 1
nonLTR- CLADE 1
nonLTR- CLADE 2
RT-nonLTR-intron
LTR
nonLTR
LTR
LTR
LTR
LTR
RT-nonLTR-intron
RT-nonLTR-intron
nonLTR
- CLADE 4

IV.

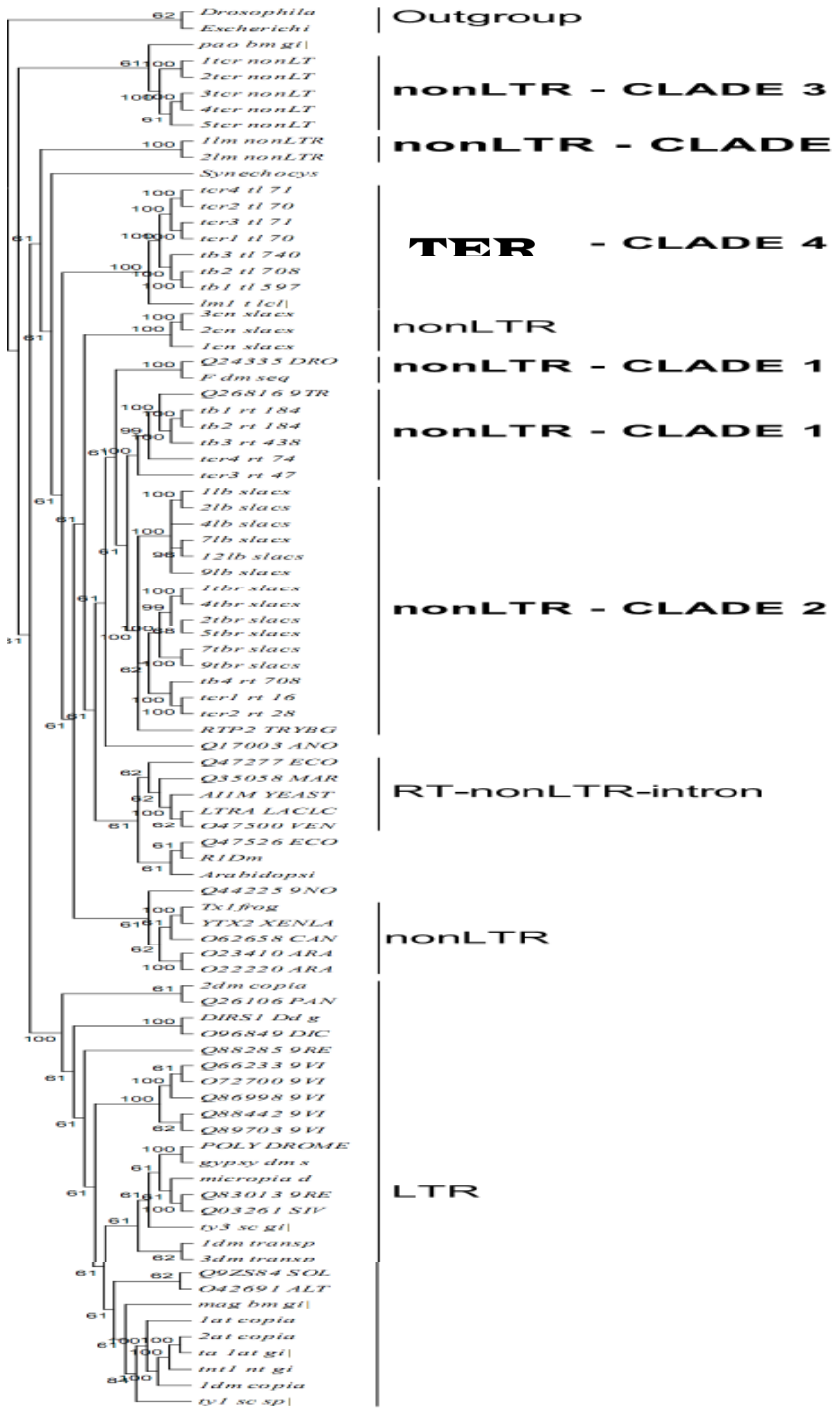


Figura 4.31 - As cinco árvores filogenéticas da transcriptases reversas em Tri-tryps obtidas com diferentes programas de filogenia

As árvores foram construídas usando cinco algoritmos filogenéticos executados com cinco programas respectivamente:

I. Algoritmo AV construído com o PAUP-AV.
II. Fluxograma de execução WEIGHBOR
III. Algoritmo MV construído com o PHYML.
IV. Algoritmo IB construído com o MRBAYES
V. Algoritmo MP construído com o PAUP-MP.

V.

4.5 ProtozoaDB: A visualização dinâmica e exploração dos genomas dos protozoários

Alguns elementos, tais como sequências e táxons, estão relacionados a estudos filogenéticos, e o GUS conta com esquemas para armazenar todas as informações extraídas das árvores filogenéticas. Portanto, o esquema Phylo possui tabelas ligadas aos outros esquemas esquemas do GUS para proporcionar a integração dos dados.

Após a geração de uma árvore filogenética, todas as informações relacionadas à execução do algoritmo são guardadas. Cada uma das informações é obtida a partir da árvore filogenética em formato nexus. Para isto, foram feitos *scripts* nas linguagens PERL e PYTHON e usadas as bibliotecas BIOPERL e BIOPYTHON.

O esquema Phylo está composto principalmente pelas seguintes entidades (figura 4.31):

(i) *TreeStudy*: (Relacionada a uma ou mais análises),

(ii) *History*: (Apresenta informações sobre as notas, alterações e atualizações do estudo. Relacionada apenas a um estudo),

(iii) *Author*: (Pode estar relacionado a muitos estudos),

(iv) *Analysis*: (Relacionado a apenas um estudo e é única no que diz respeito táxons),

(v) *Taxa*: (Relacionada com *TaxonName* (esquema GUS) armazena todas as definições científicas de um organismo e as integra à análise),

(vi) *Submission*: (É uma conexão de um determinado algoritmo - e os resultados dos seus parâmetros - à árvore),

(vii) *Tree*: (Guarda as características e a qualidade da árvore. Está relacionada a apenas uma matriz e alinhamento, mas tem um ou mais *Edge*),

(viii) *Edge*: (É o componente menor de uma árvore. Esta entidade armazena as informações em um autorrelacionamento, muito flexível para poder realizar consultas da estrutura da árvore),

(ix) *Alignment*: (Após a execução do um programa de alinhamento de sequências qualquer. Um mesmo alinhamento pode ser usado em diferentes árvores),

(x) *Matrix*: (Guarda a matriz da árvore) e

(xi) *Algorithm*: (Representa um passo do fluxograma que armazena os parâmetros de um determinado programa e as informações do próprio programa.

4.5.1 ProtozoaDB

O ProtozoaDB (http://www.biowebdb.org/protozoadb) é um banco de dados de dados desenvolvido inicialmente para guardar os dados genômicos e pós-genômicos das espécies *P. falciparum*, *E. histolytica, T. brucei*, *T. cruzi* e *L. major*. No futuro, é de se esperar a inclusão de outras espécies de protozoários assim sejam sequenciados completamente. O ProtozoaDB está baseado no banco *Genomics Unified Schema* e oferece uma moderna interface *web* para o uso amigável do usuário assim como a visualização e exploração de dados.

Este banco de dados complementa as análises bioinformáticas com ênfase no estudo das similaridades distantes usando HMM, das anotações baseadas na filogenia e das análises de ortologia. O ProtozoaDB está sendo progressivamente ligado outras bases de dados como GeneDB, PlasmoDB, TcruziDB ou TDRtargets, centrando-se na realização da combinação dinâmica de informações através de ferramentas avançadas da *web*, como *web services*.

4.5.2 Esquema Phylo

O esquema Phylo é um novo subesquema pertencente ao ProtozoaDB. Foi incorporado no GUS 3.5 a fim de armazenar os dados obtidos em experiências filogenéticas, por exemplo, a filogenia molecular usando diferentes algoritmos filogenéticos como distância, máxima parcimônia, máxima verossimilhança e inferência Bayesiana. A filogenia baseada nas análises (do ARPA) estão sendo progressivamente incorporadas no ProtozoaDB. Atualmente, os usuários podem realizar consultas das árvores filogenéticas construídas para as enzimas codificadas pelos EGM e para as sequências 18S DNAr. Nesta fase, apenas árvores obtidas por métodos de distâncias, pelo algoritmo agrupamento de vizinhos estão sendo armazenados no sistema. O objetivo final é ter todos os homólogos dos genes identificados nos cinco protozoários - *T. brucei*, *T. cruzi*, *L. major*, *E. histolytica* e *P. Falciparum* - para apoiar às inferências e armazenar a filogenia de todos os genes ortólogos e parálogos, assim como os seus domínios conservados.

As tabelas existentes no esquema Phylo foram devidamente estudadas e correlacionadas, baseadas nos *scripts* em PERL e PYTHON desenvolvidos. Foi dada ênfase a aquelas tabelas relacionadas ao armazenamento de dados das árvores filogenéticas, como mencionado na metodologia. Algumas tabelas como por exemplo tree, tree_taxa, tree_analysis e tree_edge foram modificadas, inseridas novas colunas e definidas as correlaciones entre colunas, tabelas e

esquemas por meio de chaves primárias e estrangeiras. Por outro lado, foram criadas tabelas como tree_id_taxa para fazer consistente a relação entre estas tabelas.

O processo de análise filogenética engloba o uso de diferentes ferramentas para realizar as etapas de alinhamento, geração de matrizes e criação de árvores. O acesso ordenado a essas ferramentas corresponde às etapas de um *workflow* de filogenia onde diferentes valores de parâmetros geram diferentes árvores como resultado.

4.5.3 Desenvolvimento do esquema Phylo: problemas, limitações e perspectivas

É preciso fazer um estudo abrangente sobre as tabelas que guardam as informações dos alinhamentos, modelos para a identificação de sequências por comparação (por exemplo usando HMM) - o que está sendo desenvolvido pelo nosso grupo de trabalho - e poder correlacioná-las com as tabelas do esquema Phylo que guardam as informações das ferramentas filogenéticas tal e como a versão, o algoritmo, os parâmetros, entre outros. Isto porque existe uma relação intrínseca entre as sequências identificadas incluídas no estudo, os alinhamentos e os modelos evolutivos na criação das árvores propriamente ditas. Estudos de ontologia e proveniência de dados são também necesários já que permitiram a relação direta e consistente do esquema Phylo com os outros esquemas do GUS.

O armazenamento das árvores filogenéticas em bancos de dados é uma tarefa árdua e os resultados obtidos neste trabalho marcaram os primeros intentos. No entanto, é preciso do desenvolvimento de outras tecnologias para viabilizar este objetivo, dentro delas estão: a criação de *plugins* no Phylo do GUS para a carga de dados a partir de vários arquivos no formato nexus, o acceso ao GUS por meio de sistemas de gerenciamento de *workflows*, o uso de sistemas *web* para as ferramentas filogenéticas, a sistematização das consultas no banco de dados de filogenia do GUS, o armazenamento de dados de proveniência do processo e a extensão do módulo de filogenia para a filogenômica.

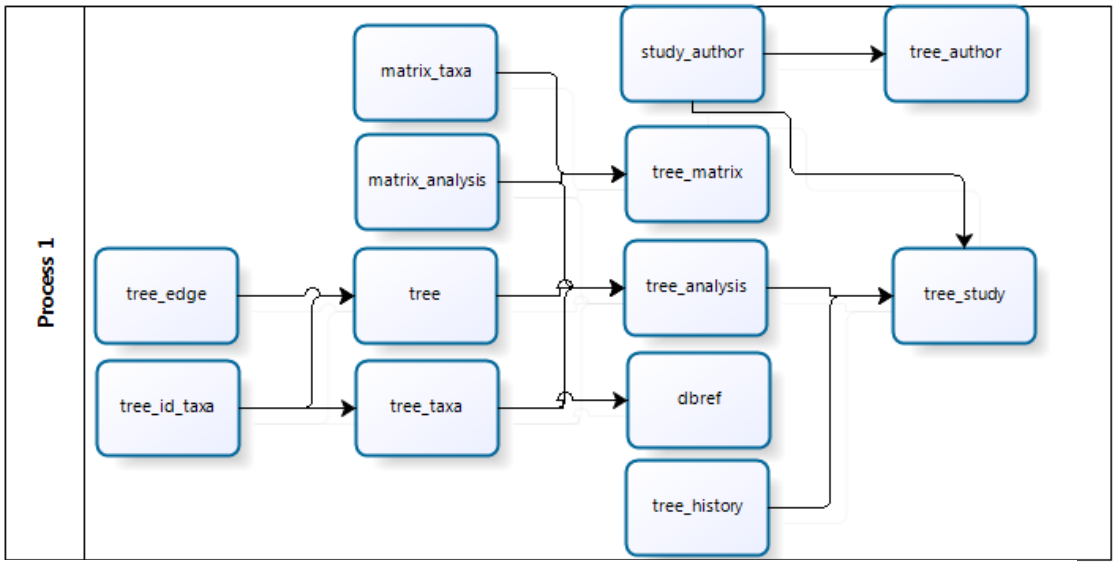


Figura 4.32 - O esquema Phylo

Tabelas, colunas e as respectivas relações do esquema Phylo

DISCUSSÃO

5.1 Sistema para a reconstrução automática de análises filogenômicas (ARPA)

Na "Era Pós Genômica", o volume de sequências nucleotídicas e protéicas incrementou-se exponencialmente, pelo qual o número de estudos filogenéticos está aumentando rapidamente, abrindo caminho às análises comparativas no nível de genomas. A filogenômica envolve o uso de genomas inteiros ou de muitas sequências para inferir filogenia de um grupo de espécies e tornou-se o padrão para reconstruir filogenias de espécies mais confiáveis (Ciccarelli *et al*. 2006; Daubin *et al*. 2002). Por outro lado, as árvores filogenômicas têm mantido a promessa de minimizar as anomalias comumente observadas na filogenia de genes individuais, devido ao uso de uma grande quantidade de dados (escala genômica) ou por serem baseadas numa quantidade grande de informação genética. Uma árvore filogenômica deveria ser o melhor reflexo da história evolutiva da espécie (Doolittle 1999; Ge *et al*. 2005). No entanto, a filogenômica pode apresentar uma série de dificuldades devido aos diferentes conjuntos de dados a serem analisados e aos diferentes métodos e parâmetros a serem utilizados, o que pode dificultar grandemente a reprodução e a comparação dos resultados. Portanto, se torna imprescindível a integração dos diferentes dados, algoritmos, programas e parâmetros em um único sistema de execução para evitar ou tentar diminuir os erros sistemáticos e estocásticos que podem terminar comprometendo a reprodutibilidade e a consistência das análises e inferências filogenéticas, filogenômicas e evolutivas. O sistema ARPA (http://arpa.biowebdb.org/) integra uma variedade de ferramentas para a análise filogenética e filogenômica. Ele foi concebido como uma resposta à crescente demanda das análises de filogenia molecular nos diferentes projetos de pesquisas. O ARPA oferece aos usuários a possibilidade de encontrar vários programas para filogenia em um ambiente *web* único e integrado, utilizando execuções automáticas.

Geralmente os *pipelines* filogenômicos tentam automatizar quatro tarefas: (i) coletar sequências de interesse e/ou homólogas, (ii) alinhá-las, (iii) gerar uma árvore ou um conjunto de árvores e (iv) analisar as informações evolutivas e filogenéticas. Para inferir filogenia em grande escala, vários *pipelines* filogenômicos automatizados têm sido desenvolvidos nos últimos anos, tal como os descritos e comparados por Fuellen em 2008. Entre estes *pipelines* podemos citar:

Pyphy (Sicheritz-Ponten *et al*. 2001), PipeAlign (Plewniak *et al*. 2003) e RiPE (Fuellen *et al*. 2005), bem como os programas PhyloGenie (Frickey *et al*. 2004), RIO (Zmasek *et al*. 2002) e FIGENIX (Gouret *et al*. 2005) e o servidor *web* Phylemon (Tarraga, 2007 # 292). Além disso, o Grupo de filogenômica da Universidade de Berkeley (http://phylogenomics.berkeley.edu) fornece um conjunto de servidores *web* para a execução dos principais programas e algoritmos filogenômicos (Glanville *et al*. 2007): (i) a enciclopédia filogenômica PhyloFacts (Krishnamurthy *et al*. 2006), (ii) o algoritmo de clusterização FlowerPower (Krishnamurthy *et al*. 2007), (iii) o algoritmo SCI-PHY (Krishnamurthy *et al*. 2007) e (iv) o servidor *web* PhyloBuilder.

Em geral, os exemplos na revisão de Fuellen (2008) demonstram que a homologia e as análises filogenômicas feitas e automatizadas em grande escala podem dar um melhor discernimento e compreensão sobre a função, distribuição e relação das sequências em determinadas espécies. No ARPA, o uso dos *scripts* na linguagem PYTHON tem facilitado diversas tarefas de estruturação e integração de arquivos e dados. O ARPA foi projetado para usuários que tenham nenhuma ou pouca experiência em filogenia, mas também pode atender às necessidades de especialistas. Os usuários básicos irão encontrar ferramentas concatenadas com parâmetros *default* em um sistema de filogenia para analisar seus dados de forma simples e robusta, enquanto os usuários avançados poderão facilmente construir e executar análises mais sofisticadas com o uso de um maior número de programas e parâmetros. Para executar a variedade de programas disponíveis no ARPA, é necessário um arquivo multifasta como entrada, para depois poder selecionar os diferentes programas e parâmetros. Em seguida, após a execução o ARPA irá apresentar os resultados na forma de diferentes arquivos de saída, incluindo o arquivo da árvore filogenética.

Existem outros sistemas muito úteis que também oferecem a opção de reconstrução filogenética. Um deles é o Phylogeny.fr (http://www.phylogeny.fr/), que oferece três modos principais: (i) *'One Click'* que utiliza os programas MUSCLE, PHYML e TREEDYN, (ii) *'Advanced'* que permite que os parâmetros de cada programa sejam personalizados pelos usuários e (iii) *'A la Carte'* que oferece mais flexibilidade e a opção, aos usuários, de poder construir seu próprio *pipeline*. Estes modos só usam um tipo de alinhamento e programa de filogenia. Por outro lado, o Phylemon é um servidor *web* que integra um conjunto de diferentes ferramentas filogenéticas e evolutivas que são executadas passo a passo com interação do usuário quem decide qual será o próximo passo (Tarraga *et al*. 2007).

O ARPA, em comparação ao Phylogeny.fr, oferece cinco programas de alinhamento - ALIGN-M, CLUSTALW, MAFFT, PROBCONS e T-COFFEE - e seis de filogenia - GARLI, MRBAYES, PAUP, PHYLIP, RAXML e WEIGHBOR - a mais, além de oferecer a opção de extrair as regiões conservadas das sequências usando dois programas diferentes - GBLOCKS e TRIMAL - tendo ainda como diferencial a execução automática e contínua dos programas. A vantagem da execução em fluxo contínuo é de customizar no início todos os programas e parâmetros, entretanto impossibilita ao usuário de fazer qualquer alteração no meio da execução. Comparado com o Phylemon, o ARPA apresenta quatro programas de alinhamento - ALIGN-M, MAFFT, PROBCONS e T-COFFEE, quatro de filogenia - GARLI, PAUP, RAXML e WEIGHBOR - e uma de extração de regiões conservadas - GBLOCKS - a mais. Por outro lado, o Phylemon disponibiliza programas do pacote PAML, que junto com outros programas e pacotes também estarão disponíveis na segunda versão do ARPA. O ARPA pode executar ao mesmo tempo seis programas de alinhamento múltiplo - ALIGN-M, CLUSTALW, MAFFT, MUSCLE, PROBCONS e T-COFFEE, dois programas de obtenção de regiões conservadas - GBLOCKS e TRIMAL - e sete programas de filogenia - GARLI, MRBAYES, PAUP, PHYLIP, PHYML, RAXML e WEIGHBOR - abrangendo os quatro algoritmos filogenéticos principais: agrupamento de vizinhos, máxima parcimônia, máxima verossimilhança e inferência Bayesiana. A execução do ARPA não precisa da interação contínua do usuário, porque ele aceita as informações de todos os parâmetros e programas antes da execução, fazendo a execução mais rápida e desta forma contornando alguns erros possíveis durante a execução. Além disso, a próxima versão do ARPA estará provida de algoritmos evolutivos e estatísticos para a comparação das árvores filogenéticas, execução em Grid dos programas e conexão com o banco de dados GUS (*Genomics Unified Schema*: www.gusdb.org) para tratar e armazenar os dados filogenéticos obtidos.

Protozoários parasitas são responsáveis por um amplo espectro de doenças em humanos e animais domésticos. Na ausência de vacinas, o uso de quimioterapia representa a melhor opção para o combate destas doenças. No entanto, o uso de quimioterápicos resultou no desenvolvimento de mecanismos de resistência nestes parasitas, limitando assim o número de medicamentos antiprotozoário que são eficazes no tratamento e controle de doenças parasitárias (Croft *et al*. 2006; Delespaux *et al*. 2007; Fidock *et al*. 2008; Gourbal *et al*. 2004; Klokouzas *et al*. 2003; Ouellette *et al*. 2001; Sobel *et al*. 1999). O conhecimento sobre os mecanismos de

resistência envolvidos podem permitir o desenvolvimento de novas drogas que minimizem ou contornem esta resistência ou mesmo que identifiquem novos alvos para o desenvolvimento de medicamentos. O presente estudo demonstrou a viabilidade do uso do ARPA na construção das árvores filogenéticas, da árvore da supermatriz e da superárvore de cinco genes relacionados à resistência a drogas em protozoários:

(i) aquaporina (AQP), transportador de membrana;

(ii) glicoproteína de 63 kDa (GP63), relacionado à virulência;

(iii) *heat-shock protein* 70 (hsp70), relacionado à proteção ao estresse;

(iv) tripanotiona redutase (TRYR), relacionado à destoxificação;

(v) MRPA, transportador de membrana.

O intuito deste estudo foi demonstrar a aplicabilidade do sistema ARPA nos dados de interesse biológico, o que foi alcançado com a inferência a partir das árvores filogenéticas e filogenômicas destes cinco genes relacionados à resistência a drogas.

Os programas que construíram as árvores filogenéticas mais consistentes foram o RAXML, o PHYML e o MRBAYES. Destes, o RAXML foi considerado o representativo e o mais consistente devido às relações filogenéticas e taxonômicas esperadas entre os táxons e aos valores consistentes de *bootstrap* representadas nas árvores. No entanto, os outros programas - PAUP-AV, PAUP-MP e WEIGHBOR - foram considerados também razoáveis para esclarecer algumas incoerências ou dúvidas nas topologias e reforçar desta maneira a inferência filogenética usando uma abordagem filogenômica. Por outro lado, dos três procedimentos para o tratamento das sequências: (i) totais, (ii) trimadas com o GBLOCK ou (iii) trimadas com o TRIMAL, os melhores resultados foram obtidos quando as sequências foram tratadas com o TRIMAL. A vantagem do uso de programas para extrair ou "trimar" os blocos conservados das sequências foi principalmente a de eliminar as regiões muito variáveis que dificultavam a inferência de filogenias robustas, e consequentemente, ao usar sequências de tamanho menor, diminuir o tempo de processamento na construção das árvores. A extração e uso de blocos conservados das sequências é principalmente vantajoso nos experimentos que: (i) possuem muitos táxons, (ii) usam programas que demandam um processamento computacional pesado (como por exemplo o MRBAYES) e/ou (iii) façam uso da abordagem filogenômica ou em grande escala. No entanto, deve ser considerado que o uso de sequências inteiras com uma similaridade alta entre elas

deveria ser a prioridade nas inferências filogenéticas. Os melhores resultados foram obtidos usando as sequências inteiras e os blocos conservados extraídos com o TRIMAL, obtendo árvores muito similares ou idênticas para os cinco genes em estudo como discutido nos seguintes parágrafos.

As árvores filogenéticas construídas com o algoritmo agrupamento de vizinhos usando todos os *hits* resultantes da comparação com o hmmpfam que pertencem a todos os táxons dos protozoários encontrados nos cinco genes de resistência a drogas em protozoários mostraram valores de *bootstrap* consistentes, isto com valores maiores de 75 nos nós basais pertencentes aos principais clados de gêneros para o gene TRYR (figura 4.2). Entretanto, isto não aconteceu nas árvores dos genes MRPA (anexo 8.3), AQP (anexo 8.4), GP63 (anexo 8.5) e hsp70 (anexo 8.6), devido ao alto número de táxons que constituíram as árvores. É conhecido que os genes MRPA (Fuellen *et al*. 2005; Klokouzas *et al*. 2003; Mukherjee *et al*. 2007; Ouellette *et al*. 2001; Sauvage *et al*. 2009) e AQP (Beitz 2005; Gourbal *et al*. 2004; Hansen *et al*. 2002; Uzcategui *et al*. 2004) são bem distribuídos e diversificados e poderia ser assumido que possam existir subtipos diferentes destes genes nas árvores respectivas.

Para o gene AQP, as árvores filogenéticas **iA**, **iB**, **iC**, **iiA**, **iiB** e **iiC** construídas usando todos os melhores *hits* - táxons dos protozoários resultantes da comparação com o hmmpfam - com o RAXML (figura 4.4.1-4.4.6) - as árvores construídas com os programas PAUP-AV, PAUP-MP, PHYML, WEIGHBOR e MRBAYES (anexo 8.8.1-8.8.15) foram consultadas também para fortalecer a inferência - mostraram que os grupos monofiléticos formados pelos gêneros *Trypanosoma*, (*bootstrap* > 99) *Leishmania* (*bootstrap* > 94), *Plasmodium* (*bootstrap* > 97) e *Entamoeba* (*bootstrap* > 95) apresentaram consistência devido aos altos valores de *bootstrap* e às relações taxonômicas da filogenia de espécies. Segundo estudos (Gourbal *et al*. 2004; Hansen *et al*. 2002; Pavlovic-Djuranovic *et al*. 2003; Uzcategui *et al*. 2004), os parasitas Apicomplexa apresentam uma redução no número de aquaporinas (AQP) em comparação aos parasitas Kinetoplastida. Dentro dos Apicomplexa, as espécies de *Plasmodium* analisadas até o momento - *P. falciparum*, *P. berghei*, *P. yoelii*, *P. chabaudi*, *P. knowlesi* - bem como *T. gondii* e *E. tenella* apresentam um único gene de aquaporina e a primeira identificação da perda de aquaporinas por um organismo eucariota foi relatada na espécie *C. parvum*. Por outro lado, até cinco genes de aquaporina foram identificados nos genomas dos Kinetoplastida *T. brucei*, *T. cruzi* e *L. major*. Um estudo comparativo feito por Beitz, em 2005 baseou-se em análises

filogenéticas dos genes da aquaporina em protozoários e indicou que as aquaporinas presentes nos gêneros *Trypanosoma* (*T. brucei*), *Plasmodium* e *Leishmania* claramente foram agrupadas no ramo das aquagliceroporinas de *E. coli* protótipo GlpF mantendo uma relação muito próxima ao grupo das aquagliceroporinas humanas 7 e 9 (Beitz 2005). As aquaporinas do filo Apicomplexa (classe Coccidia), ou seja, as aquaporinas de *T. gondii* e *E. tenella*, são muito divergentes das aquaporinas de *Plasmodium* (classe Haemosporida). As aquaporinas da classe Coccidia junto a uma série de aquaporinas não caracterizadas em *L. major* e em *T. cruzi* são as mais próximas filogeneticamente às aquaporinas específicas encontradas em *Arabidopsis thaliana* (Beitz 2005; Maurel *et al*. 1993).

Para o gene GP63, as árvores filogenéticas **iA**, **iB**, **iC**, **iiA**, **iiB** e **iiC** construídas usando todos os melhores *hits* - táxons dos protozoários resultantes da comparação com o hmmpfam - com o RAXML (figura 4.5.1-4.5.6) - as árvores construídas com os programas PAUP-AV, PAUP-MP, PHYML, WEIGHBOR e MRBAYES (anexo 8.9.1-8.9.15) foram consultadas também para fortalecer a inferência - mostraram que os grupos monofiléticos formados pelos gêneros *Trypanosoma* (*bootstrap* > 82), *Leishmania* (*bootstrap* > 98) e *Entamoeba* (*bootstrap* 100) apresentaram consistência devido aos altos valores de *bootstrap* e às relações taxonômicas da filogenia de espécies. Uma das mais notórias enzimas proteolíticas é um metaloprotease de 63 kDa (gp63) detectada em todas as espécies de *Leishmania*. Em tripanossomatídeos, as proteases mais amplamente distribuídas e altamente ativas são as enzimas proteolíticas metalo-cisteína (Branquinha *et al*. 1994; Branquinha *et al*. 1996; Mauricio *et al*. 2007; McKerrow *et al*. 1993; North *et al*. 1982). Ao longo dos últimos anos, tornou-se cada vez mais claro que as proteases produzidas por tripanossomatídeos patogênicos (especialmente *T. cruzi* e *Leishmania spp*.) desempenham um papel importante em várias etapas da infecção de hospedeiros, incluindo: absorção, penetração, sobrevivência intracelular, replicação, diferenciação, infectividade, evasão imune e nutrição (El-Sayed *et al*. 1997; Santos *et al*. 2006).

Para o gene TRYR, as árvores filogenéticas **iA**, **iB**, **iC**, **iiA**, **iiB** e **iiC** construídas usando todos os melhores *hits* - táxons dos protozoários resultantes da comparação com o hmmpfam - com o RAXML (figura 4.7.1-4.7.6) - as árvores construídas com os programas PAUP-AV, PAUP-MP, PHYML, WEIGHBOR e MRBAYES (anexo 8.11.1-8.11.15) foram consultadas também para fortalecer a inferência - mostraram que os grupos monofiléticos formados pelos gêneros *Plasmodium* (*bootstrap* > 81), *Trypanosoma* (*bootstrap* > 90), *Leishmania* (*bootstrap* >

99), *Theileria* (*bootstrap* > 99), *Cryptosporidium* (*bootstrap* > 99) e *Giardia* (*bootstrap* > 99) apresentaram consistência devido aos altos valores de *bootstrap* e às relações taxonômicas da filogenia de espécies. O TRYR é uma oxidorredutase dissulfeto dependente de NADPH que é a enzima responsável pela manutenção da tripanotiona em sua forma reduzida e acredita-se que seja o centro do sistema de defesa redox dos tripanossomatídeos (Dumas *et al*. 1997). O sistema de defesa antioxidante nos tripanossomatídeos é um potencial alvo para quimioterapia. Este sistema é baseado no tiol de baixa massa molecular "tripanotiona"; que mantêm o ambiente intracelular reduzido pela ação da enzima tripanotiona redutase (TRYR). O sistema tripanotiona é necessário para a sobrevivência do protozoário devido a que a tripanotiona ditiol é necessário para a síntese dos precursores do DNA, a homeostase do ascorbato, a desintoxicação dos hidroperóxidos e o sequestro/exportação de derivados do tiol (Singh *et al*. 2008). Este tiol está ausente no ser humano e é essencial para a sobrevivência dos parasitas. As enzimas que originam e utilizam esta molécula são alvos para o desenvolvimento de novas drogas para tratar estas doenças (Krauth-Siegel *et al*. 2003a; Krauth-Siegel *et al*. 2003b; Richardson *et al*. 2009; Schmidt *et al*. 2002). Os parasitas protozoários pertencentes à ordem Kinetoplastida contêm a tripanotiona como o seu principal tiol. Este tiol não é encontrado em outros protozoários parasitas (Dumas *et al*. 1997). Por outro lado, Ondarza em 2007, descreveu a presença de tióis e enzimas do metabolismo redox apenas em tripanossomatídeos (Singh *et al*. 2008), *Entamoeba* (Tamayo *et al*. 2005) e *Naegleria* (Ondarza 2007).

A árvore da supermatriz (anexo 8.12.1) dos genes relacionados à resistência a drogas em protozoários foi formada pelos seguintes grupos representativos: ***Paramecium***, ***Cryptosporidium***, ***Babesia***, ***Leishmania*** e ***Plasmodium***. Os melhores resultados foram obtidos com a abordagem da supermatriz devido a que a superárvore (anexo 8.12.2) formou grupos monofiléticos destes gêneros similares, mas incompletos. Estas duas abordagens foram insensíveis na formação dos principais grupos monofiléticos de protozoários esperados na topologia da árvore filogenética.

Dutilh *et al*., em 2007, comparou de forma sistemática metodologias alternativas tais como: o conteúdo de genes, os superalinhamentos, as superdistâncias (para construir uma supermatriz) e as abordagens de superárvores utilizando vários algoritmos e métodos de construção de árvore no grupo Fungi. Segundo os mesmos autores, as árvores filogenômicas reproduziram muitos dos clados de acordo com as atuais visões taxonômicas. As supermatrizes e as superárvores reproduziram a melhor filogenia de fungos que métodos fenotípicos tradicionais

(Guarro *et al*. 1999) - análises filogenéticas moleculares baseadas em RNAr (Fell *et al*. 2000; Scorzetti *et al*. 2002; Tehler *et al*. 2003) ou pequenos números de outras proteínas (Diezmann *et al*. 2004; James *et al*. 2006; Kouvelis *et al*. 2004; Kurtzman 2003), assim como também alguns estudos em grande escala (Jeffroy *et al*. 2006; Kuramae *et al*. 2006; Robbertse *et al*. 2006; Rokas *et al*. 2003; Thomarat *et al*. 2004).

5.2 Filogenômica baseada na reconstrução da árvore de espécies de protozoários

Apresenta-se neste estudo uma visão da filogenômica dos protozoários e a sua relação com outras espécies taxonômicas filogeneticamente próximas, como uma maneira de compreender melhor a posição dos protozoários na árvore de eucariotos. Estas espécies filogeneticamente próximas aos protozoários são consideradas como protozoários mitocondriais ou plastídeos, pela classificação taxonômica do banco de dados TAXONOMY (http://www.ncbi.nlm.nih.gov/taxonomy), embora pertençam a outros grupos eucarióticos. A vasta maioria da diversidade eucariótica se encontra representada por protozoários e a maior parte dos protozoários sequenciados é parasita (Chaudhary *et al*. 2005). É importante estabelecer a posição dos protozoários no interior do domínio Eukarya, embora as regras de classificação taxonômica não sejam totalmente claras (Cavalier-Smith 2005; Cavalier-Smith 1999; Levine *et al*. 1980; Turmel *et al*. 2002; Vickerman 1976). Atualmente, a maioria dos estudos filogenéticos que usam apenas um gene ou genes ribossomais não deveria ser considerada como informação completa para a inferência da filogenia de espécie. Embora, os genes RNA ribossomais e outros genes únicos tenham sido extremamente valiosos para inferir a filogenia de espécies, a filogenia de gene-único tem suas limitações (Teichmann *et al*. 1999; Wu *et al*. 2008). Atualmente, tem-se a opção de concatenar sequências de genes múltiplos com mais sinais filogenéticos para construir árvores no nível genômico, como as "árvores do genoma" - também chamadas as "árvores supermatriz". Estas árvores são menos suscetíveis a erros estocásticos relacionados ao comprimento dos genes (Daubin *et al*. 2003; Doolittle 1999; Dutilh *et al*. 2004; Jeffroy *et al*. 2006) do que as árvores construídas a partir de um único gene. Portanto, a abordagem filogenômica ao fazer uso de um número maior de genes ou regiões genômicas, permitirá a princípio, uma inferência mais confiável e representativa para a árvore de espécie.

Neste estudo foi avaliada a presença e/ou ausência do sinal filogenético

significativamente estruturado, para cada conjunto de dados, por meio da realização dos testes estatísticos. Os testes estatísticos em filogenética permitem a avaliação do grau de confiança presente em qualquer topologia a ser assumida como verdadeira. Isto foi conseguido através da realização de buscas de parcimônia exaustivas em cada conjunto de dados e comparando a distribuição assimétrica do comprimento da árvore resultante da Estatística G1 com os valores críticos publicados por Hillis e Huelsenbeck em 1992. As distribuições negativas indicam a presença de árvores mais curtas do que seria esperado se fossem geradas árvores ao acaso. Adicionalmente, o Teste *Permutation Tail Probability* (PTP) foi projetado para indicar se há qualquer sinal filogenético verdadeiro em um conjunto de dados (alinhamento de sequências) (Felsenstein 2003).

No entanto, duas dificuldades podem interferir no teste de permutação: quando existem espécies similares com a detecção de sinal filogenético e/ou quando a árvore possui um único nó interno (Felsenstein 2003). Estas dificuldades não foram observadas na presente análise. O Teste PTP (figura 4.8) tem baixo poder discriminatório por ser dependente do número de táxons e da proporção dos estados de caráter na matriz (Peres-Neto *et al*. 2000). Devido a esses motivos, foi realizada também a Estatística G1, que indica sinal filogenético nos dados utilizados na análise quando a curva é assimétrica e a árvore com maior índice de parcimônia encontra-se à esquerda desta curva. Na Estatística G1 (figura 4.9), quando uma árvore é encontrada fora da cauda da curva, em direção ao menor escore de parcimônia (menor comprimento), há a indicação de que existe sinal filogenético nos dados (Felsenstein 2003; Hillis *et al*. 1992). Por sua vez, dados que resultam em uma curva relativamente simétrica possuem pouco ou nenhum sinal filogenético.

O Teste PTP e a Estatística G1 indicaram que existe sinal filogenético nas sequências do presente, pois em ambos a curva foi assimétrica e a árvore com o menor escore posicionou-se fora da cauda da curva (Schneider 2003). Resumindo, as duas abordagens estatísticas - o Teste PTP e a Estatística G1 - demonstraram que os 31 ortólogos universais deste estudo possuem sinal filogenético. Por outro lado o uso de alinhamentos concatenados - a partir dos alinhamentos simples - ofereceu mais Caracteres Parcimoniosos Informativos em comparação ao uso do alinhamento simples por separado de um único gene.

Nas análises filogenéticas, o grupo G1 apresentou relações filogenéticas concordantes com a monofilia dos tripanossomatídeos e com as relações filogenéticas entre grupos mais próximos taxonomicamente conhecidos como diplomonas e tricomonas (Keeling *et al*. 2005;

Philippe *et al*. 2004). O gênero *Euglena* devia ter representado o grupo seguinte mais próximo das espécies da ordem Kinetoplastida, mas foi colocado incorretamente no grupo G3 devido aos poucos genes analisados. Análises filogenéticas prévias (Keeling *et al*. 2005; Philippe *et al*. 2004) confirmaram a relação muito próxima entre diplomonas (*Giardia*) e tricomonas (*Trichomonas*), e os resultados deste estudo corroboraram esta relação. As diplomonas mostraram-se estreitamente relacionadas aos Apicomplexa Alveolates como *Cryptosporidium*, aos ciliados *Tetrahymena*, *Paramecium* (estes três gêneros formaram grupos monofiléticos), *Monosiga* e *Entamoeba*. Outras Apicomplexa como *Plasmodium*, *Theileria* e *Babesia* apresentaram-se também estreitamente relacionadas. *Toxoplasma* foi erroneamente colocado no grupo G3, separado das outras Apicomplexas Alveolates, também devido ao insuficiente número de genes desta espécie encontrados nas bases de dados e utilizados neste estudo.

Cryptosporidium é um protista pertencente ao filo Apicomplexa, baseado na filogenia do RNA ribosomal, no entanto não mostrem uma relação próxima em especial aos outros membros do filo Apicomplexa (Johnson *et al*. 1990). O gênero *Cryptosporidium* é o único *Coccidium* pertencente à família Cryptosporidiidae e atualmente possue 21 espécies e aproximadamente 40 genótipos (Fayer *et al*. 2009; Xiao *et al*. 2008). *Cryptosporidium* é filogeneticamente muito distante dos Apicomplexa Haemosporidia e Coccidia (Zhu *et al*. 2000) e, dependendo do gene e método filogenético utilizado, pode ser o basal para todos os Apicomplexa examinados até agora e o grupo-irmão dos Apicomplexa gregarinas (Carreno *et al*. 1999; Leander *et al*. 2003). Por tanto, a pesar das evidências filogenéticas prévias - filogenia de genes ribossomais e individuais - (Huang *et al*. 2004; Zhu *et al*. 2000) sugere-se que *Cryptosporidium* é uma linhagem parente, mas distante dos parasitas Apicomplexa, segundo o estudo filogenômico de genes ortólogos universais realizado neste trabalho.

O reino Animalia e seus respectivos parentes unicelulares (em conjunto denominados Holozoa) mostraram uma forte afinidade com fungos (em conjunto denominados Opisthokonta) e forneceram uma forte evidência de relação com protozoários (por exemplo, *Monosiga*) (King *et al*. 2001; Philippe *et al*. 2004; Snell *et al*. 2001; Wainright *et al*. 1993). A árvore obtida no nosso estudo mostrou esta relação. A espécie *M. brevicollis* foi encontrada intimamente relacionada ao grupo G1. A posição e as relações filogenéticas dos Opisthokonta foi apoiada por muitas filogenias de genes individuais (Baldauf *et al*. 1993; Wainright *et al*. 1993) e concatenados (Baldauf *et al*. 2000; Bapteste *et al*. 2002; Philippe *et al*. 2004). A posição e as relações

filogenéticas do grupo Amoebozoa, por outro lado, também foi apoiado por várias filogenias de genes individuais e concatenados (Bapteste *et al*. 2002; Gray *et al*. 2004). Unikonta é o nome usado para a união de dois grupos individualmente bem consistentes: Amoebozoa e Opisthokonta (Cavalier-Smith 2002). Globalmente, Unikonta inclui animais e fungos, algumas amebas (por exemplo, *Entamoeba*), *slime molds* (por exemplo, *Dictyostelium*) e alguns parasitas protistas.

A filogenômica de espécies de protozoários (figura 4.10, 4.11) obtida neste trabalho mostrou uma boa concordância com estudos publicados na literatura (Baldauf *et al*. 2000; Bapteste *et al*. 2002; Chaudhary & Roos 2005; Keeling *et al*. 2005; Lang *et al*. 2002; Philippe *et al*. 2004). Infelizmente, devido ao fato de ter utilizado genomas incompletos neste estudo, a filogenia do grupo G3 não foi confiável. A explicação proposta é que as espécies que formaram este grupo G3 tinham menos de 50% (15/31) dos 31 ortólogos universais utilizados, sugerindo que o uso de menos de 15 destes genes universais, é insuficiente para obter uma inferência filogenômica robusta.

5.3 Filogenômica dos EGM em protozoários

5.3.1 Análise filogenética dos retrotransposons não-LTR

Os retrotransposons não-LTR possuem uma estrutura muito diversificada e podem ser inseridos em uma ampla variedade de diferentes tipos de alvos de DNA. Eles constituem um grupo antigo de retroelementos eucarióticos que codificam a enzima essencial transcriptase reversa, homóloga ao íntron do Grupo II e a telomerase (Nakamura *et al*. 1997; Novikova *et al*. 2009; Xiong *et al*. 1990). A maioria dos retrotransposons não-LTR indicam possuir uma distribuição não aleatória, baseado no viés observado na composição base dos sítios de inserção e na relativa especificidade de sequência da endonuclease L1 (Cost *et al*. 1998; Feng *et al*. 1996; Jurka 1997). No presente estudo, o grupo basal da filogenia dos retrotransposons em protozoários baseada na transcriptase reversa (figura 4.13) foi representado pelos não-LTR (cor dos ramos vermelha) encontrados na espécie *G. lamblia* - um dos eucariontes mais primitivos - o que sugere a sua origem ancestral e corrobora que a origem dos não-LTR coincide com a ocorrência de genomas eucarióticos (Arkhipova *et al*. 2001; Malik *et al*. 1999). Na verdade, a sua divergência e posição em relação a outros organismos eucariontes a localiza perto da transição entre procariotas e eucariotas (Hashimoto *et al*. 1995; Sogin *et al*. 1989).

Na figura 4.13, as transcriptases reversas da *G. lamblia* (círculo preto) formaram um grupo monofilético e sofreram uma duplicação que as dividiu nos subgrupos:

(1) **as pertencentes aos não-LTR e não-LTR *like*** e

(2) **as incluídas,** consideradas como não-LTR devido às relações as filogenéticas que as relacionam com outros não-LTR.

Pode ser inferido que as **transcriptases reversas *like*** e as **incluídas** pertenceriam a outros tipos de não-LTR que ainda não foram descritos ou anotados devidamente. Por outro lado, elas estão filogeneticamente mais próximas às transcriptases reversas dos retrotransposons LINE e/ou Gil. A presença dos não-LTR em *G. lamblia* pode ser inferida devido a que o elemento Gil é um dos seis elementos - Gil (ou GENIE), CRE, NeSL, R2, HERO, e R4 (Eickbush *et al*. 2002; Kojima *et al*. 2004; Kojima *et al*. 2005; Malik *et al*. 1999) - mais antigos e mais bem distribuídos na maioria dos organismos. Os retroelementos GilT e GilM são potencialmente ativos e estão representados principalmente por sequências intatas. O GilD - não encontrado neste estudo - está exclusivamente representado por exemplares inativos. Em geral, as comparações estruturais mostram que os elementos GilT e GilM possuem profundas semelhanças funcionais ao telômero associado aos retrotransposons de *Drosophila* (Burke *et al*. 2002). Arkhipova e Morrison descreveram e analisaram a paralogia encontrada entre os elementos GilT e GilM, assim como a relação estreita entre o GilM e o LINE (Arkhipova & Morrison 2001), o que foi observado neste trabalho.

Com base na filogenia apresentada na figura 4.13, infere-se que os retrotransposons não-LTR SLACS foram introduzidos na família *Trypanosomatidae* e fixados ao longo da evolução, adaptando-se aos eventos de especiação e espalhando-se em *Leishmania* e *Trypanosoma*. Curiosamente os SLACS foram encontrados apenas em *L. brasiliensis* como os SLACS-*like* e não foram encontrados em outras espécies de *Leishmania*, reafirmado por análises previas (Peacock *et al*. 2007). Em *Trypanosoma*, os SLACS foram identificados em *T. brucei* (Aksoy *et al*. 1990) em quanto os retrotransposons não-LTR CZAR em *T. cruzi* (Gabriel *et al*. 1990). Como reportado em outros estudos, o genoma do *T. brucei* apresenta duas famílias de não-LTR - o ingi (Kimmel *et al*. 1987) e o SLACS (Aksoy *et al*. 1987; Carrington *et al*. 1987) - que se integram exclusivamente nos mini-éxons (SL). Homólogos dos SLACS têm sido identificados em todas as subespécies de *T. brucei* e muito mais recentemente reportado em *L. brasiliensis* (Peacock *et al*. 2007). Os retroelementos SL RNA-específicos têm sido caracterizados nos parasitas protozoários,

como *T. cruzi* (CZAR) (Gabriel *et al*. 1990) e *C. fasciulata* (CRE1) (Villanueva *et al*. 1991), assim como também em *Caenorhabditis elegans* (Ne-SL-1) (Malik *et al*. 2000b). *Leishmania* poderia ter perdido alguma maquinaria relacionada à geração da diversidade, comparado com outros eucariotos, devido à pressão seletiva nos rearranjos cromossomais, A falta de elementos retrotransponíveis favoreceria a estabilidade do cromossomo, o que pode ser observado nos genomas de *L. major* e *L. infantum*. Diferentes classes de elementos retrotransponíveis - os não-LTR *ingi*/L1*Tc* e SLACS/CZAR, e o LTR VIPER - estão presentes nos parasitas Kinetoplastida *T. brucei e T. cruzi*, porém o genoma de *L. major* só tem remanescentes de elementos relacionados ao *ingi*/L1T*c* DIREs, sugerindo sua perda durante a evolução nas linhagens de *Leishmania*. Ainda, *L. infantum* e *L. braziliensis* também contêm os retroelementos *ingi*/L1*Tc* DIREs (Peacock *et al*. 2007).

5.3.2 Análise filogenética dos retrotransposons LTR

O grupo das transcriptases reversas pertencentes aos LTR (cor de ramos verde) apresentou monofilia e poderia ser considerado como polifilético ao grupo não-LTR. Os LTR foram representados pelos retroelementos Copia e Gypsy (figura 4.13). Os Copia encontraram-se bem adaptados no gênero *Phytophthora* - *P. infestans* e *P. parasítica*, dado sua significativa distribuição. Já os Gypsy estiveram representados pelos tipos Ty3-gypsy em *R. salina* que foi o mais representativo e melhor distribuído e gypsy-*like* em *P. yezoensis*. *P. sojae* acumulou as classes Gypsy Ps-1 e Gypsy Ps-2. *E. tenella* foi incluído neste grupo baseado nas relações filogenéticas, mas a sua transcriptase reversa foi encontrada sem nenhuma especificação sobre o tipo de retroelemento ao qual pertence.

Espécies de *Phytophthora* são conhecidas por apresentar grande variação fenotípica, tanto no campo e na cultura como na reprodução assexuada. A base genética desse fenômeno não é clara, mas poderia ser devido à instabilidade do genoma, talvez causada por elementos transponíveis, conversão gênica e/ou recombinação mitótica. Sequências similares aos elementos transponíveis são abundantes em genomas de *Phytophthora*. Por exemplo, sequências com similaridade aos retrotransposons das classes Copia e Gypsy/Ty têm sido descritos. As sequências Gypsy-*like* foram detectadas em 29 espécies de *Phytophthora* e variam em abundância de 10 a 10.000 cópias por genoma (W. Morgan e S. Kamoun, dados não publicados). As primeiras famílias dos elementos curtos intercalados relacionados aos SINE em oomycetes

foram identificadas pela exploração das sequências genômicas de DNA em *P. infestans*. Quinze famílias relacionadas aos tRNA-SINE foram identificadas, no entanto, sete elementos foram demonstrados ser exclusivos de *P. infestans* (Whisson *et al*. 2005).

5.3.3 Análise filogenética da telomerase

O grupo mais representativo das telomerases (cor de ramos verdes) foi encontrado nas espécies da família *Trypanosomatidae*. Este grupo monofilético separou-se em dois clados, um deles representado pelo gênero *Trypanosoma* - *T. brucei* e *T. cruzi* - e o outro pelo gênero *Leishmania* - *L. infantum*, *L. donovani*, *L. amazonensis*, *L. brasiliensis* e *L. major*. Os outros grupos monofiléticos independentes presentes na árvore foram formados pelas telomerases das espécies *P. tetraurelia*, *T. thermophila*, *C. hominis*, *C. parvum*, *G. lamblia*, *P. yoelii yoelii* e *T. annulata*, o que indica que mesmo sendo uma proteína conservada, a telomerase apresenta uma relação intra-gênero mais estreita.

Um estudo de identificação, caracterização e supressão da telomerase foi descrita em *T. brucei* (Dreesen *et al*. 2005), onde ortológos da telomerase foram identificados nas sequências do genoma de *T. cruzi* (49,1% de similaridade com a telomerase de *T. brucei*) e *L. major* (33,6% de similaridade) (Dreesen *et al*. 2005). Por outro lado, uma análise filogenética (Giardini *et al*. 2006) sugere que as telomerases de *Leishmania* estiveram mais relacionadas filogeneticamente entre elas do que a outras telomerases de *Trypanosoma* e de outros eucariotos (Giardini *et al*. 2006).

A identificação de uma telomerase em *G. lamblia* semelhante à encontrada na maioria dos outros eucariotos sugere que a telomerase remonta ao primeiro marcador existente da evolução eucariótica (Malik *et al*. 2000a). Com a perspectiva de utilizar a telomerase como alvo para antimaláricos, esta enzima tem sido caracterizada e analisada em *Plasmodium* (Figueiredo *et al*. 2005a; Figueiredo *et al*. 2005b). Para futuros testes de drogas anti-telomerase em modelos experimentais da malária (ratos e macacos), será útil saber quais sequências da telomerase das espécies de *Plasmodium* causam a malária nestes animais. Na espécie *P. falciparum*, vários homólogos hipotéticos de proteínas teloméricas específicas foram identificados e estão sendo caracterizados (Figueiredo & Scherf 2005a; Figueiredo *et al*. 2005b). A atividade da telomerase foi detectada em extratos nucleares semipurificados dos estágios de *P. falciparum* em sangue. Foi relatada a identificação e a caracterização de uma telomerase hipotética no parasita da malária

humana *P. falciparum* e dos genes da telomerase nas espécies *P. yoelii*, *P. knowlesi* e *P. berghei* (Figueiredo & Scherf 2005a; Figueiredo *et al*. 2005b).

5.3.4 Análise filogenética das proteínas gag, gag-pol e pol, da integrase e da ribonuclease H

A árvore filogenética da proteína gag (figura 4.14) foi restrita nas espécies ***D. discoideum***, ***P. tricornotum***, ***T. pseudonana*** e ***P. infestans***. As árvores reconstruídas com os algoritmos agrupamento de vizinhos, máxima parcimônia e inferência Bayesiana mostraram as topologias mais consistentes. As quatro sequências presentes em *D. discoideum* mostraram uma monofilia com um valor de *bootstrap* de 100. Elas foram anotadas como retrotransposons LTR Skipper por Leng *et al*., em 1998. A árvore filogenética da figura 4.15 indica que a proteína gag-pol foi encontrada em dois grupos monofiléticos de retrotransposons Ty1/copia-*like* representados pelas espécies, ***P. tricornotum*** e ***T. pseudonana***. A proteína pol (figura 4.16) foi encontrada em ***P. tricornotum*** pertencendo ao retrotransposon Ty1/copia-*like*, em ***T. pseudonana*** pertencendo ao Ty1/copia-*like* e Ty3/gypsy-*like*) e ***P. infestans*** pertencendo ao retrotransposon gypsy-*like* GypsyPi-1a.

A árvore da integrase (figura 4.17) apresentou três grupos monofiléticos formado por 21 sequências em ***T. thermophila***, 34 em ***T. vaginalis*** e duas em ***P. yoelii yoelii*** (Carlton *et al*. 2002). O ciliado *T. thermophila* é um dos organismos modelo unicelular eucariótico usado na biologia e seu estudo tem contribuído para novas descobertas biológicas fundamentais, tais como RNA catalítico, repetições teloméricas, telomerase e a função da acetilação de histonas (Eisen *et al*. 2006). Em *T. thermophila*, várias famílias de retransposon-*like* foram identificadas. A família Tel-1 é restrita ao micronúcleo e estruturalmente se assemelham aos transposons (Cherry *et al*. 1985). A sequência do genoma do protista *T. vaginalis* - patógeno sexualmente transmissível em humanos - cujo tamanho aproximado é de 160 Mb, reflete a recente expansão maciça do material genético. As 59 famílias das repetições mais comuns identificadas constituem no conjunto aproximadamente 39 Mb do genoma e podem ser classificados como (i) virus-*like*, (ii) transposon-*like*, incluindo aproximadamente 1000 cópias do elemento *mariner* identificado pela primeira vez fora dos animais, (iii) retrotransposon-*like* e (iv) repetições não classificadas (Carlton *et al*. 2007).

A árvore da ribonuclease H (figura 4.18) apresentou três principais classes: ribonuclease

H, ribonuclease HI e ribonuclease HII. Estas classes foram encontradas unicamente nos gêneros *Leishmania* e *Trypanosoma*. Mishra *et al*., (2005) sustentou que a ribonuclease H foi distribuída e diferenciada nos grupos de bactérias e eucariotas/arquea. Além disso, Arudchandran *et al*., (2002) antecipou esta afirmação e concluiu que pelo menos um gene codificando a proteína ribonuclease H-*like* está presente em todos os genomas de procariotas e arqueas. Na maioria das vezes existem dois genes, ou uma combinação entre as ribonuclease: HI e HII ou HII e HIII. Por outro lado, pouco se sabe sobre o número e tipos de ribonucleases H em eucariontes. A diversificação da ribonuclease H assim como a incapacidade de encontrar mais de uma ribonuclease H em alguns organismos podem ser um reflexo da falta de algoritmos de similaridade sensíveis na procura de sequências de proteínas homólogas. Mishra *et al*., (2005) identificaram quatro genes distintos de ribonucleases H em *Leishmania*: uma ribonuclease H do tipo I - LRNase HI - e três do tipo II - LRNase HIIA, LRNase HIIB e LRNase HIIC. Este estudo não só revelou a importância da ribonuclease HIIC LRNase na biologia dos Kinetoplastida, mas também identificou um potencial alvo molecular para a quimioterapia anti-leishmaniose.

5.3.5 Estratégia para a detecção *in silico* da seleção positiva de genes dos EGM em protozoários

Os três modelos que permitem sítios com seleção positiva - M2, M3 e M8 - tiveram a razão d_N/d_S (ω_2) maior que 1 e foram estatisticamente significativos para a o teste LRT para todos os genes de EGM ($P \leq 0.05$). No entanto, o modelo M3 resultou ser extrema e estatisticamente significativo para a o teste LRT ($P \leq 0.0001$) (tabela 4.9-4.16). De acordo com Yang e Nielsen (2002) o LRT entre os modelos M0 e M3 é mais um teste de variabilidade da razão ω entre os sítios do que um teste de seleção propriamente. O próprio manual do pacote PAML aconselha a utilização de dois pares de modelos (M1-M2 e M7-M8) para detecção da seleção positiva. Assim, o método de sítio detectou a seleção positiva nos dois pares de modelos propostos (M1-M2 e M7-M8), porém indica uma alta variabilidade da razão ω entre os sítios (M0-M3). Estes resultados demonstraram a importância da detecção de seleção positiva no nível de sítios ou códons individuais, em vez de analisar o gene em sua totalidade, tal como demonstrado na literatura (Suzuki *et al*. 2001a; Suzuki *et al*. 2001b; Yang *et al*. 2000) (Fitch *et al*. 1997).

Existem poucos estudos de seleção molecular feitos em EGM. Um estudo de seleção foi feito na transcriptase reversa pertencente à superfamília dos retrotransposons LTR Ty1-copia em

várias espécies de mono e dicotiledôneas (Navarro-Quezada *et al*. 2002), bem como em genomas das gramíneas (Matsuoka *et al*. 1999) que indicaram a presença de seleção diversificadora. Baucom *et al*., em 2009 analisaram os padrões de variação genética entre os retrotransposons LTR pertencentes a espécie *Oryza sativa* e descobriram evidências de seleção purificadora forte em todas as regiões do gene, mas também indícios de que raros episódios de seleção positiva e adaptação ocorreram no genoma do hospedeiro. Além disso, os resultados de Baucom *et al*., (2009) indicam que os retrotransposons LTR exibem diferentes, mas previsíveis padrões de variação nas sequências, dependendo da data da sua transposição, sugerindo que os retrotransposons LTR, independentemente da superfamília e classificações da família, mostram similares histórias de vida evolutivas. Um estudo em *S. cerevisiae* feito por Sawyer e Malik em 2006 também confirmou a presença de seleção positiva atuando nos genes chamados de NHEJ, que teria sido adquirida após estes genes interagirem com os retrotransposons LTR Ty (Sawyer *et al*. 2006).

Como podem estes resultados díspares ser conciliados? Um modelo de evolução de sequência no qual a seleção purificadora usando códons predomina, mas onde raros episódios de seleção positiva ocorrem, permite a expansão populacional e uma explosão transposicional (Baucom *et al*. 2009), o que poderia resolver todos os conflitos aparentes nos resultados de ambas as hipóteses. Neste cenário, um regime de seleção purificadora severa seria a regra, no entanto, raramente, um retrotransposon LTR mutante não reconhecido pelo genoma hospedeiro poderia surgir e não seria silenciado, pelo menos, enquanto a população de retrotransposons LTR permanece em baixa frequência. Esta hipótese suporta a idéia de que regimes episódicos de seleção positiva e adaptação do genoma ocorrem, seguido da distribuição dos retroelementos no genoma hospedeiro e períodos subsequentes de inatividade.

O fenômeno da seleção positiva nos EGM nos protozoários pode ter surgido como resultado de diferentes processos adaptativos, por exemplo, a adaptação destes para otimizar o processo de infecção, evasão de resposta imune do hospedeiro, adaptação ao meio ou nicho biológico e/ou a diversificação funcional (Anisimova *et al*. 2007a; Anisimova *et al*. 2007b).

5.4 A telomerase e os elementos de retrotransposição em Tri-tryps baseados em análises filogenéticas dos domínios da transcriptase reversa

As topologias das árvores construídas com os programas PAUP-AV, WEIGHBOR, PAUP-MP e MRBAYES (figura 4.29-4.31) foram muito similares e foram consideradas como as mais consistentes. As árvores construídas com o **PAUP-AV** e o **MRBAYES** apresentaram um ramo mais externo que formou dois grupos parafiléticos mais próximos filogeneticamente, o **Clado 3** (não-LTR) e o **Clado 4** (telomerases). Por outro lado, estes clados mudaram de posição, sendo o mais externo o Clado 4 para a árvore do WEIGHBOR e o Clado 3 para a árvore do PAUP-MP. Para todas as árvores, o Clado 4 foi formado por sequências de telomerases pertencentes aos Tri-tryps que estão mais próximas a um grupo de retrotransposons não-LTR nas espécies *L. major* e *T. cruzi* pertencentes ao Clado 3. O Clado 4 mostrou a espécie *L. major* (lm) como o táxon mais externo. Um grupo interno parafilético foi formado por *T. brucei* (tb ou tbr) e *T. cruzi* (tcr). A árvore do PHYML não determinou a monofilia para o Clado 4 e apresentou as espécies sem relação alguma, isto é como uma polifilia aparente. Esta incoerência poderia ser explicada pela possível falta de consistência do PHYML na reconstrução da filogenia de sequências altamente divergentes ou consideradas homólogas distantes como a transcriptase reversa.

As árvores construídas com os programas PAUP-AV, WEIGHBOR, PAUP-MP e MRBAYES mostraram diferenças no subgrupo de *T. brucei*. As árvores do WEIGHBOR e do PAUP-MP apresentaram um isolado de *T. brucei* (tb3) como o táxon mais externo, enquanto outros isolados *T. brucei* (tb1) e *T. brucei* (tb2) seriam subgrupos obtidos após possíveis duplicações. Por outro lado, as árvores do PAUP-AV e do MRBAYES sustentariam esta hipótese, ou também a possibilidade da perda de uma quarta sequência que se apresentaria como formada após de um evento de duplicação com *T. brucei* (tb3). Estes quatro algoritmos sustentam a parafilia em *T. cruzi* com a presença de uma duplicação recente validada pelo valor de *bootstrap* de 100. Além disso, estas árvores mostraram parafilia entre o **Clado 1** (não-LTR) e o **Clado 2** (não-LTR SLACS, ORF, sequências hipotéticas). O Clado 2 contém sequências de SLACS em *L. brasiliensis*, *T. brucei*, *T. brucei gambiense* e *T. cruzi*.

Estudos filogenéticos prévios realizados por Eickbush (Eickbush 1994; Eickbush 1997; Eickbush & Malik 2002) utilizaram os sete motivos conservados em 80 transcriptases reversas - incluindo às telomerases. Estas telomerases foram posicionadas e incluídas no grupo dos retrotransposons não-LTR, quando a árvore está enraizada com a RNA polimerase. Acredita-se que os ramos da telomerase espalharam-se ancestralmente entre o grupo que contém os não-LTR, os elementos bacterianos e as organelas. Isto sugere que a telomerase em eucariotas ancestrais poderia ter dado origem aos retroelementos parasitas e esta hipótese sugere ainda que os não-LTR seriam os mais antigos entre os retroelementos parasitas móveis deste ramo (Eickbush 1994). Por outro lado, Eickbush em 1997, mostrou a filogenia das transcriptases reversas dos procariotos e elementos derivados das organelas (Eickbush 1997). A implicação do enraizamento desta árvore é que o grupo formado pelas transcriptases reversas dos não-LTR seria o mais antigo, e que a telomerase e os LTR divergiram desta linhagem. Esta interpretação sugere também a importância da função celular da manutenção dos telômeros em eucariotos primitivos ou ancestrais a qual foi cumprida pelo recrutamento de um gene de transcriptase reversa a partir de um elemento móvel parasita (Eickbush 1997). O principal argumento em favor desta hipótese é que ela não exige a transferência de sequências de eucariotos a procariotos (Eickbush 1994; Eickbush 1997).

Embora os domínios da telomerase são evolutivamente relacionados com outras transcriptases reversas, a determinação da distância evolutiva ou topologia exata entre a telomerase e os retroelementos - ou mesmo entre retroelementos - é um imprecisa. Os retroelementos acumulam mutações muito mais rápido quando estão duplicados através do ciclo de cDNA pela transcriptase reversa, em relação a quando estão integrados em cromossomos e replicados pela polimerase DNA cromossômica. A telomerase não é um elemento transponível e, portanto, nunca deve ser esperado que seja o responsável da replicação devido à baixa fidelidade da transcriptase reversa. No entanto a inclusão de elementos altamente transponíveis ou vírus que passam por um ciclo de cDNA com frequência pode obscurecer a topologia da árvore filogenética (Nakamura *et al*. 1997). Algumas dificuldades adicionais na análise derivam do pequeno número de aminoácidos dos domínios da transcriptase reversa - apenas 178 aminoácidos podem ser alinhados com razoável confiança - e da relativamente baixa similaridade nas sequências entre os membros das espécies que conformam a árvore. Existem evidências de que a árvore da telomerase sofre com esses problemas (Nakamura *et al*. 1997). Por outro lado, a

situação pode manter as expectativas de poder inferir filogenia utilizando os sete domínios da transcriptase reversa dos elementos, pois mesmo tendo uma baixa conservação entre as suas sequências são formados clados independentes que diferenciam os diferentes subgrupos da transcriptase reversa pertencentes aos retrotransposons, tal e como foi demonstrado neste estudo.

5.5 ProtozoaDB: A visualização dinâmica e exploração dos genomas dos protozoários

A arquitetura do ProtozoaDB (http://protozoadb.biowebdb.org) permite o armazenamento e a integração flexível de dados genômicos em um repositório central. Esta base de dados não se destina a duplicar os esforços de outros bancos de dados semelhantes, como o GeneDB (http://www.genedb.org), o PlasmoDB (http://www.plasmodb.org), o TcruziDB (http://www.tcruzidb.org) ou mesmo o TDRtargets (http://www.tdrtargets.org), pelo contrário destina-se a complementar as informações genômicas através de análises baseadas em similaridades distantes usando perfis HMM, em análises filogenéticas e em análises de ortologia.

O ProtozoaDB integra várias aplicações da bioinformática: (i) bases de dados heterogêneas, (ii) diferentes sistemas de anotação, (iii) diferentes ferramentas de análise e (iv) computação distribuída; as quais permitem realizar análises comparativas de vários protozoários ao mesmo tempo em uma única interface. Além disso, o ProtozoaDB possui *links* que permitem a conexão com os outros bancos de dados semelhantes, e porém com as informações que os conformam.

O diferencial do ProtozoaDB, além da centralização dos dados genômicos e pós-genômicos, é a nova idéia de exploração e visualização dos genomas por meio das metodologias baseadas na filogenia e na comparação por similaridade, assim como também as que estão sendo implementadas baseadas em ontologia e *druggability*. Isto permitirá fazer um estudo comparativo mais abrangente, rápido e completo, e desta maneira obter uma inferência biológica muito mais consistente e, porém inovadora.

Considerando tanto os dados genômicos existentes de importantes parasitas protozoários humanos, bem como os dados de várias outras espécies patogênicas relevantes - *P. vivax*, *T. gondii*, *C. parvum*, *T. rangeli*, *L. braziliensis* e *L. chagasi* (Liolios *et al*. 2006) - o ProtozoaDB está sendo desenvolvido para hospedar inicialmente dados genômicos e pós-genômicos de cinco protozoários, *P. falciparum*, *E. histolytica*, *T. brucei*, *T. cruzi* e *L. major*, e devido que este possui

o esquema baseado no GUS, poderá ser feita a integração e a comparação direta entre os genomas destes microorganismos, ou mesmo até, com outros diferentes genomas de eucariotos. Desde o ponto de vista computacional, isto racionalizaria o custo de tempo e energia, por outro lado desde o ponto de vista biológico, este tipo de estudo pode fazer a inferência biológica, filogenética e evolutiva mais consistente e fácil, especialmente para os usuários biólogos sem muita experiência computacional.

O novo esquema chamado de Phylo - concebido e incorporado no GUS - armazena os dados obtidos de experiências filogenéticas que podem usar diferentes algoritmos filogenéticos como distância, máxima parcimônia, máxima verossimilhança ou inferência Bayesiana. O Phylo tem um esquema muito mais abrangente devido a que também podem ser guardadas e relacionadas outras informações como as correspondentes ao alinhamento, à matriz filogenética ou ao algoritmo filogenético usado, e permitirá, além disso, incorporar no futuro estudos de proveniência de dados. Por outro lado, o esquema do Phylo também relaciona os táxons da árvore ao TAXONOMY e adicionalmente armazena os táxons como entidades independentes e relacionadas - com *scripts* em PERL, o que permite poder navegar pelos táxons da árvore, já que estes táxons mantiveram as correlações de herança. Esta diferencia faz com que o usuário possa realizar consultas biológicas e navegar entre os táxons de uma ou de várias árvores.

A criação de bancos de dados especialmente projetados para executar consultas complexas sobre as informações contidas nas árvores filogenéticas é uma tarefa bastante custosa e muito importante, porém pouco explorada, para o estudo da filogenia. Armazenar a árvore filogenética de maneira flexível em um banco de dados corresponde a relacionar cada registro a um ramo da árvore, ou seja, aos táxons. Desta maneira, é possível aumentar a complexidade das consultas e analisar cada ramo da árvore ou de várias árvores, em uma só consulta.

O TreeBASE (http://www.treebase.org/treebase-web/home.html) é atualmente o banco de dados mais completo que armazena as informações filogenéticas dos organismos publicados. Ele armazena as informações - os táxons da árvore e a matriz - que provêem da árvore em formato NEXUS e relaciona cada um dos táxons ao banco de dados TAXONOMY (Chen *et al*. 2008; Page 2007).

CAPÍTULO 6 - CONCLUSÕES

1. O desenvolvimento e implementação do sistema ARPA se mostrou como uma alternativa viável, eficiente, fácil e de tempo reduzido para as análises filogenômicas de sequências.

2. Os programas que construíram as árvores filogenéticas mais consistentes foram o RAXML, o PHYML e o MRBAYES. Destes, o RAXML foi considerado o melhor.

3. As árvores construídas usando as sequências inteiras e as trimadas com o TRIMAL apresentaram os melhores resultados.

4. A árvore filogenética do gene TRYR construída com o algoritmo agrupamento de vizinhos usando todos os *hits* dos Genes de Resistência a Drogas em Protozoários foi a única que mostrou valores de *bootstrap* consistentes.

5. A filogenia do gene AQP apresentou os gêneros *Trypanosoma*, *Leishmania*, *Plasmodium* e *Entamoeba* formando grupos monofiléticos representativos. O gene aquaporina pertencente aos protozoários é considerado um alvo potencial para a quimioterapia.

6. A filogenia do gene GP63 demonstrou que os gêneros *Trypanosoma* e *Leishmania* formaram grupos monofiléticos representativos. Torna-se cada vez mais claro que as proteases produzidas por tripanossomatídeos patogênicos (especialmente *T. cruzi* e *Leishmania spp*.) desempenham um papel importante em várias etapas da infecção de hospedeiros.

7. A filogenia do gene TRYR: os gêneros *Plasmodium*, *Trypanosoma* e *Leishmania* formaram grupos monofiléticos representativos. As enzimas que originam e utilizam a tripanotiona ditiol pertencente ao sistema tripanotiona são alvos para o desenvolvimento de novas drogas para tratar as doenças geradas por estes protozoários.

8. A supermatriz foi formada pelos seguintes grupos representativos: ***Paramecium***, ***Cryptosporidium***, ***Babesia***, ***Leishmania*** e ***Plasmodium***. Os melhores resultados foram obtidos com a abordagem da supermatriz. A superárvore formou grupos monofiléticos destes gêneros similares, mas incompletos.

9. Foi apresentada uma visão geral baseada na filogenômica para a reconstrução da árvore de espécies dos protozoários. As relações entre os grupos de protozoários estão de acordo com estudos anteriores, os quais determinaram uma monofilia para este grupo.

10. As informações filogenéticas inferidas a partir do clado G3 da árvore de espécies dos protozoários são questionáveis, devido aos poucos genes encontrados, sugerindo que o uso de menos de 15 ortólogos universais para a reconstrução filogenômica não é confiável. A inclusão de mais dados/genes é necessária para obter uma árvore robusta.

11. A metodologia baseada na filogenômica usando a abordagem da supermatriz provou ser confiável para os dados dos genomas dos protozoários, indicando que quanto mais ortólogos universais utilizados, melhor é a análise; e que utilizando as sequências inteiras dos ortólogos universais, ou apenas um bloco conservado delas, pode-se obter resultados confiáveis similares.

12. Nas reconstruções dos genes dos EGM, o RAXML foi o programa mais consistente ao lidar com os diferentes níveis de polimorfismos destes genes.

13. O modelo M3 resultou ser extrema e estatisticamente significativo para a o teste LRT ($P \leq 0.0001$) e indicou uma alta variabilidade da razão ω entre os sítios. Os modelos M7 e M8 resultaram ser estatisticamente significativos e indicaram a presença de seleção positiva para todos os genes dos EGM.

14. Na filogenia da transcriptase e da telomerase de Tri-tryps, observou-se que esta formou um grupo monofilético isolado e mais relacionado ao grupo da transcriptase pertencente aos retrotransposons não-LTR.

15. O esquema Phylo concebido e incorporado no GUS 3.5 armazena os dados obtidos de experiências filogenéticas, mantendo as relações de herança filogenética entre cada um dos táxons, o que permite não só reconstruir a árvore como também realizar consultas usando as informações dos ramos, nós e táxons da árvore.

CAPÍTULO 7 - REFERÊNCIAS BIBLIOGRÁFICAS

Abrahamsen MS, Templeton TJ, Enomoto S *et al*. (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum Science* 304, 441-5.

Adams MD, Celniker SE, Holt RA *et al*. (2000) The genome sequence of *Drosophila melanogaster Science* 287, 2185-95.

Agrafiotis DK, Bandyopadhyay D, Wegner JK, Vlijmen H (2007) Recent advances in chemoinformatics *J Chem Inf Model* 47, 1279-93.

Ajioka JW, Fitzpatrick JM, Reitter CP (2001) *Toxoplasma gondii* genomics: shedding light on pathogenesis and chemotherapy *Expert Rev Mol Med* 2001, 1-19.

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy *Genetics* 136, 927-35.

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA *Genetics* 139, 1067-76.

Aksoy S, Lalor TM, Martin J, Van der Ploeg LH, Richards FF (1987) Multiple copies of a retroposon interrupt spliced leader RNA genes in the African trypanosome, *Trypanosoma gambiense Embo J* 6, 3819-26.

Aksoy S, Williams S, Chang S, Richards FF (1990) SLACS retrotransposon from *Trypanosoma brucei gambiense* is similar to mammalian LINEs *Nucleic Acids Res* 18, 785-92.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool *J Mol Biol* 215, 403-10.

Anisimova M, Bielawski J, Dunn K, Yang Z (2007a) Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes *BMC Evol Biol* 7, 154.

Anisimova M, Yang Z (2007b) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites *Mol Biol Evol* 24, 1219-28.

Anzai T, Takahashi H, Fujiwara H (2001) Sequence-specific recognition and cleavage of telomeric repeat (TTAGG)(n) by endonuclease of non-long terminal repeat retrotransposon TRAS1 *Mol Cell Biol* 21, 100-8.

Arango E, Carmona-Fonseca J, Blair S (2008) [In vitro susceptibility of Colombian *Plasmodium falciparum* isolates to different antimalarial drugs] *Biomedica* 28, 213-23.

Arkhipova IR, Morrison HG (2001) Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead *Proc Natl Acad Sci U S A* 98, 14497-502.

Arevalo J, Ramirez L, Adaui V *et al*. (2007) Influence of *Leishmania* (Viannia) species on the response to antimonial treatment in patients with American tegumentary leishmaniasis *J Infect Dis* 195, 1846-51.

Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB (2003) Retroelements containing introns in diverse invertebrate taxa *Nat Genet* 33, 123-4.

Ashutosh, Sundar S, Goyal N (2007) Molecular mechanisms of antimony resistance in *Leishmania J Med Microbiol* 56, 143-53.

Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-Prot: juggling between evolution and stability *Brief Bioinform* 5, 39-55.

Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins *Proc Natl Acad Sci U S A* 90, 11558-62.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data *Science* 290, 972-7.

Bansal D, Sehgal R, Chawla Y, Mahajan RC, Malla N (2004) In vitro activity of antiamoebic drugs against clinical isolates of *Entamoeba histolytica* and *Entamoeba dispar Ann Clin Microbiol Antimicrob* 3, 27.

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes *Genome Res* 12, 1269-76.

Bapteste E, Brinkmann H, Lee JA *et al*. (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba Proc Natl Acad Sci U S A* 99, 1414-9.

Barrett MP, Burchmore RJ, Stich A *et al*. (2003) The trypanosomiases *Lancet* 362, 1469-80.

Barrett MP, Gilbert IH (2002) Perspectives for new drugs against trypanosomiasis and leishmaniasis *Curr Top Med Chem* 2, 471-82.

Barton GJ, Sternberg MJ (1987) Evaluation and improvements in the automatic alignment of protein sequences *Protein Eng* 1, 89-94.

Bateman A, Coin L, Durbin R *et al*. (2004) The Pfam protein families database *Nucleic Acids Res* 32, D138-41.

Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL (2009) Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome *Genome Res* 19, 243-54.

Beitz E (2005) Aquaporins from pathogenic protozoan parasites: structure, function and potential for chemotherapy *Biol Cell* 97, 373-83.

Belfort M, Derbyshire V, Parker M, Cousineau B, Lambowitz A (2002) Mobile introns: Pathways and proteins. In ea eds. N.L. Craig ed. *In Mobile DNA II*; pp. 761–83. Washington, D.C: American Society for Microbiology.

Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution *Plant Mol Biol* 42, 251-69.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2004) GenBank: update *Nucleic Acids Res* 32, D23-6.

Berezikov E, Bucheton A, Busseau I (2000) A search for reverse transcriptase-coding sequences reveals new non-LTR retrotransposons in the genome of *Drosophila melanogaster Genome Biol* 1, RESEARCH0012.

Berriman M, Ghedin E, Hertz-Fowler C *et al*. (2005) The genome of the African trypanosome *Trypanosoma brucei Science* 309, 416-22.

Biessmann H, Mason JM, Ferry K *et al*. (1990) Addition of telomere-associated HeT DNA sequences "heals" broken chromosome ends in *Drosophila Cell* 61, 663-73.

Boeckmann B, Bairoch A, Apweiler R *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003 *Nucleic Acids Res* 31, 365-70.

Boeke JD (1997) LINEs and Alus--the polyA connection *Nat Genet* 16, 6-7.

Brandonisio O, Spinelli R (2002) Immune response to parasitic infections--an introduction *Curr Drug Targets Immune Endocr Metabol Disord* 2, 193-9.

Branquinha MH, Vermelho AB, Goldenberg S, Bonaldo MC (1994) Characterization of proteinases in trypanosomatids *Braz J Med Biol Res* 27, 495-9.

Branquinha MH, Vermelho AB, Goldenberg S, Bonaldo MC (1996) Ubiquity of cysteine- and metalloproteinase activities in a wide range of trypanosomatids *J Eukaryot Microbiol* 43, 131-5.

Bringaud F, Biteau N, Melville SE *et al*. (2002a) A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of *Trypanosoma brucei Eukaryot Cell* 1, 137-51.

Bringaud F, Biteau N, Zuiderwijk E *et al*. (2004) The ingi and RIME non-LTR retrotransposons are not randomly distributed in the genome of *Trypanosoma brucei Mol Biol Evol* 21, 520-8.

Bringaud F, Garcia-Perez JL, Heras SR *et al*. (2002b) Identification of non-autonomous non-LTR retrotransposons in the genome of *Trypanosoma cruzi Mol Biochem Parasitol* 124, 73-8.

Brosius J, Tiedge H (1995) Reverse transcriptase: mediator of genomic plasticity *Virus Genes* 11, 163-79.

Brown D, Sjolander K (2006) Functional classification using phylogenomic inference *PLoS Comput Biol* 2, e77.

Bruno WJ, Socci ND, Halpern AL (2000) Weighted neighbor joining: a likelihood-based

approach to distance-based phylogeny reconstruction *Mol Biol Evol* 17, 189-97.

Burke WD, Malik HS, Rich SM, Eickbush TH (2002) Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, *Giardia lamblia Mol Biol Evol* 19, 619-30.

Burri C, Keiser J (2001) Pharmacokinetic investigations in patients from northern Angola refractory to melarsoprol treatment *Trop Med Int Health* 6, 412-20.

Campos RM, Nascimento M, Ferraz JC *et al*. (2008) Distinct mitochondrial HSP70 homologues conserved in various *Leishmania* species suggest novel biological functions *Mol Biochem Parasitol* 160, 157-62.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses *Bioinformatics*.

Carlton JM, Angiuoli SV, Suh BB *et al*. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii Nature* 419, 512-9.

Carlton JM, Hirt RP, Silva JC *et al*. (2007) Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis Science* 315, 207-12.

Carreno RA, Martin DS, Barta JR (1999). *Cryptosporidium* is more closely related to the gregarines than to coccidia as shown by phylogenetic analysis of apicomplexan parasites inferred using small-subunit ribosomal RNA gene sequences. *Parasitol Res* 85(11):899-904.

Carrington M, Roditi I, Williams RO (1987) The structure and transcription of an element interspersed between tandem arrays of mini-exon donor RNA genes in *Trypanosoma brucei Nucleic Acids Res* 15, 10179-98.

Carstens BC, Bankhead A, 3rd, Joyce P, Sullivan J (2005) Testing population genetic structure using parametric bootstrapping and MIGRATE-N *Genetica* 124, 71-5.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis *Mol Biol Evol* 17, 540-52.

Cavalier-Smith T (1999) Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree *J Eukaryot Microbiol* 46, 347-66.

Cavalier-Smith T (2002) The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa *Int J Syst Evol Microbiol* 52, 297-354.

Cavalier-Smith T (2005) Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion *Ann Bot (Lond)* 95, 147-75.

Chapados BR, Chai Q, Hosfield DJ, Qiu J, Shen B, Tainer JA (2001) Structural biochemistry of a type 2 RNase H: RNA primer recognition and removal during DNA replication *J Mol Biol* 307, 541-56.

Chapman KB, Bystrom AS, Boeke JD (1992) Initiator methionine tRNA is essential for Ty1 transposition in yeast *Proc Natl Acad Sci U S A* 89, 3236-40.

Chaudhary K, Roos DS (2005) Protozoan genomics for drug discovery *Nat Biotechnol* 23, 1089-91.

Chen D, Burleigh JG, Bansal MS, Fernandez-Baca D (2008) PhyloFinder: an intelligent search engine for phylogenetic tree databases *BMC Evol Biol* 8, 90.

Cherry JM, Blackburn EH (1985) The internally located telomeric sequences in the germ-line chromosomes of *Tetrahymena* are at the ends of transposon-like elements *Cell* 43, 747-58.

Chiurillo MA, Cano I, Da Silveira JF, Ramirez JL (1999) Organization of telomeric and sub-telomeric regions of chromosomes from the protozoan parasite *Trypanosoma cruzi Mol Biochem Parasitol* 100, 173-83.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life *Science* 311, 1283-7.

Citerne HL, Luo D, Pennington RT, Coen E, Cronk QC (2003) A phylogenomic investigation of CYCLOIDEA-like TCP genes in the Leguminosae *Plant Physiol* 131, 1042-53.

Coffin M, Stephen H (1997) *Retroviruses*: Cold Spring Harbor Laboratory Press.

Conte MG, Gaillard S, Droc G, Perin C (2008) Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants *BMC Genomics* 9, 183.

Cost GJ, Boeke JD (1998) Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure *Biochemistry* 37, 18081-93.

Covey SN (1986) Amino acid sequence homology in gag region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus *Nucleic Acids Res* 14, 623-33.

Craig NL, Craigie R, Gellert M, Lambowitz AM (2002) *Mobile DNA II* Washington, DC: American Society for Microbiology (ASM Press).

Creevey CJ, McInerney JO (2005) Clann: investigating phylogenetic information through supertree analyses *Bioinformatics* 21, 390-2.

Croft SL, Seifert K, Yardley V (2006a) Current scenario of drug development for leishmaniasis *Indian J Med Res* 123, 399-410.

Croft SL, Sundar S, Fairlamb AH (2006b) Drug resistance in leishmaniasis *Clin Microbiol Rev* 19, 111-26.
Dagan T, Martin W (2006) The tree of one percent *Genome Biol* 7, 118.

Dai L, Zimmerly S (2002a) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior *Nucleic Acids Res* 30, 1091-102.

Dai L, Zimmerly S (2002b) The dispersal of five group II introns among natural populations of *Escherichia coli Rna* 8, 1294-307.

Darnell JE, Doolittle WF (1986) Speculations on the early course of evolution *Proc Natl Acad Sci U S A* 83, 1271-5.

Daubin V, Gouy M, Perriere G (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history *Genome Res* 12, 1080-90.

Daubin V, Moran NA, Ochman H (2003) Phylogenetics and the cohesion of bacterial genomes *Science* 301, 829-32.

Davila AM, Lorenzini DM, Mendes PN *et al*. (2005) GARSA: genomic analysis resources for sequence annotation *Bioinformatics* 21, 4302-3.

Davila AM, Majiwa PA, Grisard EC, Aksoy S, Melville SE (2003) Comparative genomics to uncover the secrets of tsetse and livestock-infective trypanosomes *Trends Parasitol* 19, 436-9.

De Souza W (2002) Basic cell biology of *Trypanosoma cruzi Curr Pharm Des* 8, 269-85.

Decuypere S, Rijal S, Yardley V *et al*. (2005) Gene expression analysis of the mechanism of natural Sb(V) resistance in *Leishmania donovani* isolates from Nepal *Antimicrob Agents Chemother* 49, 4616-21.

Delespaux V, de Koning HP (2007) Drugs and drug resistance in African trypanosomiasis *Drug Resist Updat* 10, 30-50.

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life *Nat Rev Genet* 6, 361-75.

Deppenmeier U, Johann A, Hartsch T *et al*. (2002) The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea *J Mol Microbiol Biotechnol* 4, 453-61.

Diezmann S, Cox CJ, Schonian G, Vilgalys RJ, Mitchell TG (2004) Phylogeny and evolution of medical species of *Candida* and related taxa: a multigenic analysis *J Clin Microbiol* 42, 5624-35.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment *Genome Res* 15, 330-40.

Doak TG, Doerder FP, Jahn CL, Herrick G (1994) A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif *Proc Natl Acad Sci U S A* 91, 942-6.

Dobzhansky T (1973) Nothing in Biology Makes Sense Except in the Light of Evolution *The American Biology Teacher* 35, 125-9.

Doolittle RF, Feng DF, Johnson MS, McClure MA (1989) Origins and evolutionary relationships of retroviruses *Q Rev Biol* 64, 1-30.

Doolittle WF (1999) Phylogenetic classification and the universal tree *Science* 284, 2124-9.

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution *Nature* 284, 601-3.

Dopazo H, Santoyo J, Dopazo J (2004) Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species *Bioinformatics* 20 Suppl 1, i116-21.

Dreesen O, Li B, Cross GA (2005) Telomere structure and shortening in telomerase-deficient *Trypanosoma brucei Nucleic Acids Res* 33, 4536-43.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees *BMC Evol Biol* 7, 214.

Dumas C, Ouellette M, Tovar J *et al*. (1997) Disruption of the trypanothione reductase gene of *Leishmania* decreases its ability to survive oxidative stress in macrophages *Embo J* 16, 2590-8.

Dutilh BE, Huynen MA, Bruno WJ, Snel B (2004) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise *J Mol Evol* 58, 527-39.

Dutilh BE, van Noort V, van der Heijden RT, Boekhout T, Snel B, Huynen MA (2007) Assessment of phylogenomic and orthology approaches for phylogenetic inference *Bioinformatics* 23, 815-24.

Eddy SR (1998) Profile hidden Markov models *Bioinformatics* 14, 755-63.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Res* 32, 1792-7.

Edgar RC, Sjolander K (2003a) SATCHMO: sequence alignment and tree construction using hidden Markov models *Bioinformatics* 19, 1404-11.

Edgar RC, Sjolander K (2003b) Simultaneous sequence alignment and tree construction using hidden Markov models *Pac Symp Biocomput*, 180-91.

Eickbush TH (1992) Transposing without ends: the non-LTR retrotransposable elements *New Biol* 4, 430-40.

Eickbush TH (1994) Origin and evolutionary relationship of retroelements. In SS Morse ed. *The evolutionary biology of viruses*; pp. 121–57. Rockefeller University, New York, NY, USA.: Raven Press.

Eickbush TH (1997) Telomerase and retrotransposons: which came first? *Science* 277, 911-2.

Eickbush TH (1999) Mobile introns: retrohoming by complete reverse splicing *Curr Biol* 9,

R11-4.

Eickbush TH, Malik HS (2002) Origins and Evolution of Retrotransposons. In N Craig, Craigie, R., Gellert, M., and Lambowitz, A ed. *Mobile DNA II*; pp. 1111-44. Washington, DC: ASM Press.

Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis *Genome Res* 8, 163-7.

Eisen JA, Hanawalt PC (1999) A phylogenomic study of DNA repair genes, proteins, and processes *Mutat Res* 435, 171-213.

Eisen JA, Wu M (2002) Phylogenetic analysis and gene functional predictions: phylogenomics in action *Theor Popul Biol* 61, 481-7.

Eisen JA, Fraser CM (2003) Phylogenomics: intersection of evolution and genomics *Science* 300, 1706-7.

Eisen JA, Coyne RS, Wu M *et al*. (2006) Macronuclear genome sequence of the ciliate *Tetrahymena* thermophila, a model eukaryote *PLoS Biol* 4, e286.

El-Sayed NM, Donelson JE (1997) African trypanosomes have differentially expressed genes encoding homologues of the *Leishmania* GP63 surface protease *J Biol Chem* 272, 26742-8.

El-Sayed NM, Ghedin E, Song J *et al*. (2003) The sequence and analysis of *Trypanosoma brucei* chromosome II *Nucleic Acids Res* 31, 4856-63.

El-Sayed NM, Myler PJ, Bartholomeu DC *et al*. (2005a) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease *Science* 309, 409-15.

El-Sayed NM, Myler PJ, Blandin G *et al*. (2005b) Comparative genomics of trypanosomatid parasitic protozoa *Science* 309, 404-9.

Engel A (2000) Structural analyses of the aquaporin super-family *Nephrol Dial Transplant* 15 Suppl 6, 23-5.

Faith DP, Cranston PS (1991) Could a cladograma this short have arisen by chance alone? On permutation tests for cladistic structure. *Cladistics* 7, 1-28.

Fajkus J, Sykorova E, Leitch AR (2005) Telomeres in evolution and evolution of telomeres *Chromosome Res* 13, 469-79.

Fawcett DH, Lister CK, Kellett E, Finnegan DJ (1986) Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs *Cell* 47, 1007-15.

Fayer R, Santin M (2009) *Cryptosporidium xiaoi* n. sp. (Apicomplexa: Cryptosporidiidae) in sheep (Ovis aries). *Vet Parasitol* 164(2-4):192-200.

Fell JW, Boekhout T, Fonseca A, Scorzetti G, Statzell-Tallman A (2000) Biodiversity and systematics of basidiomycetous yeasts as determined by large-subunit rDNA D1/D2 domain sequence analysis *Int J Syst Evol Microbiol* 50 Pt 3, 1351-71.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach *J Mol Evol* 17, 368-76.

Felsenstein J (1989) *PHYLIP— PHYLogeny Inference Package*.

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods *Methods Enzymol* 266, 418-27.

Felsenstein J (2003) *Inferring phylogenies*: Sinauer, Sunderland, MA, USA.

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) Version 3.6. Distributed by the Author. Department of Genome Sciences, University of Washington, Seattle.

Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition *Cell* 87, 905-16.

Fidock DA, Eastman RT, Ward SA, Meshnick SR (2008) Recent highlights in antimalarial drug resistance and chemotherapy research *Trends Parasitol* 24, 537-44.

Figueiredo L, Scherf A (2005a) *Plasmodium* telomeres and telomerase: the usual actors in an unusual scenario *Chromosome Res* 13, 517-24.

Figueiredo LM, Rocha EP, Mancio-Silva L, Prevost C, Hernandez-Verdun D, Scherf A (2005b) The unusually large *Plasmodium* telomerase reverse-transcriptase localizes in a discrete compartment associated with the nucleolus *Nucleic Acids Res* 33, 1111-22.

Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution *Trends Genet* 5, 103-7.

Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A *Proc Natl Acad Sci U S A* 94, 7712-8.
Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: a test and review of evidence *Am Nat* 160, 712-26.

Frickey T, Lupas AN (2004) PhyloGenie: automated phylome generation and analysis *Nucleic Acids Res* 32, 5231-8.

Fuellen G, Spitzer M, Cullen P, Lorkowski S (2005) Correspondence of function and phylogeny of ABC proteins based on an automated analysis of 20 model protein data sets *Proteins* 61, 888-99.

Fuellen G (2008) Homology and phylogeny and their automated inference *Naturwissenschaften* 95, 469-81.

Gabriel A, Yen TJ, Schwartz DC *et al*. (1990) A rapidly rearranging retrotransposon within the

miniexon gene locus of *Crithidia fasciculata Mol Cell Biol* 10, 615-24.

Gadelle D, Filee J, Buhler C, Forterre P (2003) Phylogenomics of type II DNA topoisomerases *Bioessays* 25, 232-42.

Gardner MJ, Hall N, Fung E *et al*. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum Nature* 419, 498-511.

Ge F, Wang LS, Kim J (2005) The cobweb of life revealed by genome-scale estimates of horizontal gene transfer *PLoS Biol* 3, e316.

Ghedin E, Bringaud F, Peterson J *et al*. (2004) Gene synteny and evolution of genome architecture in trypanosomatids *Mol Biochem Parasitol* 134, 183-91.

Giardini MA, Lira CB, Conte FF *et al*. (2006) The putative telomerase reverse transcriptase component of *Leishmania amazonensis*: gene cloning and characterization *Parasitol Res* 98, 447-54.

Gilbert W, Glynias M (1993) On the ancient nature of introns *Gene* 135, 137-44.

Gilbert D (2003) Sequence file format conversion with command-line readseq *Curr Protoc Bioinformatics* Appendix 1, Appendix 1E.

Gillespie J (1992) *The Causes of Molecular Evolution.*: Oxford University Press.

Glanville JG, Kirshner D, Krishnamurthy N, Sjolander K (2007) Berkeley Phylogenomics Group web servers: resources for structural phylogenomic analysis *Nucleic Acids Res* 35, W27-32.

Glauser A, Braun R (1994) TUBIS, a fossilized retroposon in the tubulin gene cluster of *Trypanosoma brucei Biochim Biophys Acta* 1218, 99-101.

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences *Mol Biol Evol* 11, 725-36.
Gotoh O (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments *J Mol Biol* 264, 823-38.

Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure *J Mol Biol* 313, 903-19.

Gourbal B, Sonuc N, Bhattacharjee H *et al*. (2004) Drug uptake and modulation of drug resistance in *Leishmania* by an aquaglyceroporin *J Biol Chem* 279, 31010-7.

Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EG (2005) FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform *BMC Bioinformatics* 6, 198.

Graham WV, Tcheng DK, Shirk AL, Attene-Ramos MS, Welge ME, Gaskins HR (2004) Phylomat: an automated protein motif analysis tool for phylogenomics *J Proteome Res* 3, 1289-91.

Gray MW, Lang BF, Burger G (2004) Mitochondria of protists *Annu Rev Genet* 38, 477-524.

Greider CW, Blackburn EH (1987) The telomere terminal transferase of *Tetrahymena* is a ribonucleoprotein enzyme with two kinds of primer specificity *Cell* 51, 887-98.

Guarro J, GeneJ, Stchigel AM (1999) Developments in fungal taxonomy *Clin Microbiol Rev* 12, 454-500.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood *Syst Biol* 52, 696-704.

Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference *Nucleic Acids Res* 33, W557-9.

Hall N, Berriman M, Lennard NJ *et al*. (2003) The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism *Nucleic Acids Res* 31, 4864-73.

Hansen M, Kun JF, Schultz JE, Beitz E (2002) A single, bi-functional aquaglyceroporin in blood-stage *Plasmodium falciparum* malaria parasites *J Biol Chem* 277, 4874-82.

Harrington L, Zhou W, McPhail T *et al*. (1997) Human telomerase contains evolutionarily conserved catalytic and structural subunits *Genes Dev* 11, 3109-15.

Hasan G, Turner MJ, Cordingley JS (1984) Ribosomal RNA genes of *Trypanosoma brucei*: mapping the regions specifying the six small ribosomal RNAs *Gene* 27, 75-86.

Hashimoto T, Nakamura Y, Kamaishi T *et al*. (1995) Phylogenetic place of mitochondrion-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2 *Mol Biol Evol* 12, 782-93.

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks *Proc Natl Acad Sci U S A* 89, 10915-9.

Herskowitz IH, Muller HJ (1954) Evidence against a Straight End-to-End Alignment of Chromosomes in *Drosophila* Spermatozoa *Genetics* 39, 836-50.
Hertz-Fowler C, Berriman M (2004) Continuing tsetse and *Trypanosoma* genome sequencing projects *Trends Parasitol* 20, 308-9; author reply 9-10.

Hillis DM, Huelsenbeck JP (1992) Signal, noise, and reliability in molecular phylogenetic analyses *J Hered* 83, 189-95.

Hoare CA (1966) The classification of mammalian trypanosomes *Ergeb Mikrobiol Immunitatsforsch Exp Ther* 39, 43-57.

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches *Nat Rev Genet* 4, 275-84.

Holmes I (2002) Transcendent elements: whole-genome transposon screens and open evolutionary questions *Genome Res* 12, 1152-5.

Huang J, Mullapudi N, Sicheritz-Ponten T, Kissinger JC (2004) A first glimpse into the pattern and scale of gene transfer in Apicomplexa *Int J Parasitol* 34, 265-74.

Huang J, Mullapudi N, Lancto CA *et al* (2004) Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol* 5(11):R88.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees *Bioinformatics* 17, 754-5.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T (2007) The human phylome *Genome Biol* 8, R109.

Hughey R, Krogh A (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method *Comput Appl Biosci* 12, 95-107.

Hulo N, Sigrist CJ, Le Saux V *et al*. (2004) Recent improvements to the PROSITE database *Nucleic Acids Res* 32, D134-7.

Ivens AC, Peacock CS, Worthey EA *et al*. (2005) The genome of the kinetoplastid parasite, *Leishmania major Science* 309, 436-42.

James TY, Kauff F, Schoch CL *et al*. (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny *Nature* 443, 818-22.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? *Trends Genet* 22, 225-31.

Jiang YW (2002) Transcriptional cosuppression of yeast Ty1 retrotransposons *Genes Dev* 16, 467-78.

Johnson AM, Fielke R, Lumb R *et al*. (1990) Phylogenetic relationships of *Cryptosporidium* determined by ribosomal RNA sequence comparison. *Int J Parasitol*; 20(2):141-7.

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons *Proc Natl Acad Sci U S A* 94, 1872-7.

Kajikawa M, Okada N (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence *Cell* 111, 433-44.

Katoh K, Kuma K, Miyata T, Toh H (2005) Improvement in the accuracy of multiple sequence

alignment program MAFFT *Genome Inform* 16, 22-33.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform *Nucleic Acids Res* 30, 3059-66.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program *Brief Bioinform* 9, 286-98.

Kashiwagi T, Jeanteur D, Haruki M, Katayanagi K, Kanaya S, Morikawa K (1996) Proposal for new catalytic roles for two invariant residues in *Escherichia coli* ribonuclease HI *Protein Eng* 9, 857-67.

Katinka MD, Duprat S, Cornillot E *et al*. (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi Nature* 414, 450-3.

Kazazian HH, Jr. (2004) Mobile elements: drivers of genome evolution *Science* 303, 1626-32.

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McLnerney JO (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified *BMC Evol Biol* 6, 29.

Keeling PJ, Burger G, Durnford DG *et al*. (2005) The tree of eukaryotes *Trends Ecol Evol* 20, 670-6.

Keeling PJ, Doolittle WF (1996) A non-canonical genetic code in an early diverging eukaryotic lineage *Embo J* 15, 2285-90.

Kempken F, Jacobsen S, Kuck U (1998) Distribution of the fungal transposon Restless: full-length and truncated copies in closely related strains *Fungal Genet Biol* 25, 110-8.

Ketting RF, Haverkamp TH, van Luenen HG, Plasterk RH (1999) Mut-7 of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD *Cell* 99, 133-41.

Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution *Evolution* 55, 1-24.

Kilian A, Bowtell DD, Abud HE *et al*. (1997) Isolation of a candidate human telomerase catalytic subunit gene, which reveals complex splicing patterns in different cell types *Hum Mol Genet* 6, 2011-9.

Kimmel BE, ole-MoiYoi OK, Young JR (1987) Ingi, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs *Mol Cell Biol* 7, 1465-75.

Kimura M (1989) The neutral theory of molecular evolution and the world view of the neutralists *Genome* 31, 24-31.

King N, Carroll SB (2001) A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution *Proc Natl Acad Sci U S A* 98, 15032-7.

Klenk HP, Spitzer M, Ochsenreiter T, Fuellen G (2004) Phylogenomics of hyperthermophilic Archaea and Bacteria *Biochem Soc Trans* 32, 175-8.

Klokouzas A, Shahi S, Hladky SB, Barrand MA, van Veen HW (2003) ABC transporters and drug resistance in parasitic protozoa *Int J Antimicrob Agents* 22, 301-17.

Kojima KK, Fujiwara H (2004) Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets *Mol Biol Evol* 21, 207-17.

Kojima KK, Fujiwara H (2005) An extraordinary retrotransposon family encoding dual endonucleases *Genome Res* 15, 1106-17.

Kouvelis VN, Ghikas DV, Typas MA (2004) The analysis of the complete mitochondrial genome of *Lecanicillium muscarium* (synonym *Verticillium lecanii*) suggests a minimum common gene organization in mtDNAs of Sordariomycetes: phylogenetic implications *Fungal Genet Biol* 41, 930-40.

Krauth-Siegel RL, Inhoff O (2003a) Parasite-specific trypanothione reductase as a drug target molecule *Parasitol Res* 90 Suppl 2, S77-85.

Krauth-Siegel RL, Meiering SK, Schmidt H (2003b) The parasite-specific trypanothione metabolism of *Trypanosoma* and *Leishmania Biol Chem* 384, 539-49.
Krishnamurthy N, Brown D, Sjolander K (2007) FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function *BMC Evol Biol* 7 Suppl 1, S12.

Krishnamurthy N, Brown DP, Kirshner D, Sjolander K (2006) PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification *Genome Biol* 7, R83.

Krishnamurthy N, Sjolander K (2005) Phylogenomic inference of protein molecular function *Curr Protoc Bioinformatics* Chapter 6, Unit 6 9.

Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters *Bioinformatics* 22, 768-70.

Kuiper MT, Lambowitz AM (1988) A novel reverse transcriptase activity associated with mitochondrial plasmids of *Neurospora Cell* 55, 693-704.

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences *Brief Bioinform* 9, 299-306.

Kuramae EE, Robert V, Snel B, Weiss M, Boekhout T (2006) Phylogenomics reveal a robust fungal tree of life *FEMS Yeast Res* 6, 1213-20.

Kurtzman CP (2003) Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora FEMS Yeast Res* 4, 233-45.

Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G (2002) The closest unicellular relatives of animals *Curr Biol* 12, 1773-8.

Larkin MA, Blackshields G, Brown NP *et al*. (2007) Clustal W and Clustal X version 2.0 *Bioinformatics* 23, 2947-8.

Leander BS, Clopton RE, Keeling PJ (2003) Phylogeny of gregarines (Apicomplexa) as inferred from small-subunit rDNA and beta-tubulin. *Int J Syst Evol Microbiol* 53(Pt 1):345-54.

Levin HL (1995) A novel mechanism of self-primed reverse transcription defines a new family of retroelements *Mol Cell Biol* 15, 3310-7.

Levine ND, Corliss JO, Cox FE *et al*. (1980) A newly revised classification of the protozoa *J Protozool* 27, 37-58.

Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM (1993) Transposons in place of telomeric repeats at a *Drosophila* telomere *Cell* 75, 1083-93.

Lingner J, Cech TR, Hughes TR, Lundblad V (1997a) Three Ever Shorter Telomere (EST) genes are dispensable for in vitro yeast telomerase activity *Proc Natl Acad Sci U S A* 94, 11190-5.

Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR (1997b) Reverse transcriptase motifs in the catalytic subunit of telomerase *Science* 276, 561-7.

Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide *Nucleic Acids Res* 34, D332-4.

Logsdon JM, Jr. (1998) The recent origins of spliceosomal introns revisited *Curr Opin Genet Dev* 8, 637-48.

Lohe AR, De Aguiar D, Hartl DL (1997) Mutations in the mariner transposase: the D,D(35)E consensus sequence is nonfunctional *Proc Natl Acad Sci U S A* 94, 1293-7.

Louis EJ (2002) Are *Drosophila* telomeres an exception or the rule? *Genome Biol* 3, REVIEWS0007.

Madera M, Gough J (2002) A comparison of profile hidden Markov model procedures for remote homology detection *Nucleic Acids Res* 30, 4321-8.

Malik HS, Burke WD, Eickbush TH (1999) The age and evolution of non-LTR retrotransposable elements *Mol Biol Evol* 16, 793-805.
Malik HS, Burke WD, Eickbush TH (2000a) Putative telomerase catalytic subunits from *Giardia lamblia* and *Caenorhabditis elegans Gene* 251, 101-8.

Malik HS, Eickbush TH (2000b) NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans Genetics* 154, 193-203.

Malik HS, Eickbush TH (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses *Genome Res* 11, 1187-97.

Marcotte EM, Xenarios I, van Der Bliek AM, Eisenberg D (2000) Localizing proteins in the cell from their phylogenetic profiles *Proc Natl Acad Sci U S A* 97, 12115-20.

Margulis L (1996) Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life *Proc Natl Acad Sci U S A* 93, 1071-6.

Martin F, Maranon C, Olivares M, Alonso C, Lopez MC (1995) Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi*: homology of the first ORF with the ape family of DNA repair enzymes *J Mol Biol* 247, 49-59.

Matsuoka Y, Tsunewaki K (1999) Evolutionary dynamics of Ty1-copia group retrotransposons in grass shown by reverse transcriptase domain analysis *Mol Biol Evol* 16, 208-17.

Maurel C, Reizer J, Schroeder JI, Chrispeels MJ (1993) The vacuolar membrane protein gamma-TIP creates water specific channels in *Xenopus oocytes Embo J* 12, 2241-7.

Mauricio IL, Gaunt MW, Stothard JR, Miles MA (2007) Glycoprotein 63 (gp63) genes show gene conversion and reveal the evolution of Old World *Leishmania Int J Parasitol* 37, 565-76.

McClintock B (1938) The Production of Homozygous Deficient Tissues with Mutant Characteristics by Means of the Aberrant Mitotic Behavior of Ring-Shaped Chromosomes *Genetics* 23, 315-76.

McClintock B (1941a) The Association of Mutants with Homozygous Deficiencies in *Zea Mays Genetics* 26, 542-71.

McClintock B (1941b) The Stability of Broken Ends of Chromosomes in *Zea Mays Genetics* 26, 234-82.

McClintock B (1984) The significance of responses of the genome to challenge *Science* 226, 792-801.

McClure MA, Richardson HS, Clinton RA, Hepp CM, Crowther BA, Donaldson EF (2005) Automated characterization of potentially active retroid agents in the human genome *Genomics* 85, 512-23.

McKerrow JH, Sun E, Rosenthal PJ, Bouvier J (1993) The proteases and pathogenicity of parasitic protozoa *Annu Rev Microbiol* 47, 821-53.

Meyerson M, Counter CM, Eaton EN *et al*. (1997) hEST2, the putative human telomerase catalytic subunit gene, is up-regulated in tumor cells and during immortalization *Cell* 90, 785-95.

Misra S, Bennett J, Friew YN *et al*. (2005) A type II ribonuclease H from *Leishmania* mitochondria: an enzyme essential for the growth of the parasite *Mol Biochem Parasitol* 143, 135-45.

Miyata T, Yasunaga T (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application *J Mol Evol* 16, 23-36.

Mukherjee A, Padmanabhan PK, Singh S *et al*. (2007) Role of ABC transporter MRPA, gamma-glutamylcysteine synthetase and ornithine decarboxylase in natural antimony-resistant isolates of *Leishmania donovani J Antimicrob Chemother* 59, 204-11.

Munoz DP, Collins K (2004) Biochemical properties of *Trypanosoma cruzi* telomerase *Nucleic Acids Res* 32, 5214-22.

Murphy NB, Pays A, Tebabi P *et al*. (1987) *Trypanosoma brucei* repeated element with unusual structural and transcriptional properties *J Mol Biol* 195, 855-71.

Myler PJ (2008) Searching the Tritryp genomes for drug targets *Adv Exp Med Biol* 625, 133-40.

Nakamura TM, Morin GB, Chapman KB *et al*. (1997) Telomerase catalytic subunit homologs from fission yeast and human *Science* 277, 955-9.
Nakamura TM, Cech TR (1998a) Reversing time: origin of telomerase *Cell* 92, 587-90.

Nakamura TM, Cooper JP, Cech TR (1998b) Two modes of survival of fission yeast without telomerase *Science* 282, 493-6.

Nakamura TM, Morin GB, Chapman KB *et al*. (1997) Telomerase catalytic subunit homologs from fission yeast and human *Science* 277, 955-9.

Nakayama J, Saito M, Nakamura H, Matsuura A, Ishikawa F (1997) TLP1: a gene encoding a protein component of mammalian telomerase is a novel member of WD repeats family *Cell* 88, 875-84.

Navarro-Quezada A, Schoen DJ (2002) Sequence evolution and copy number of Ty1-copia retrotransposons in diverse plant genomes *Proc Natl Acad Sci U S A* 99, 268-73.

Neuwald AF, Poleksic A (2000) PSI-BLAST searches using hidden markov models of structural repeats: prediction of an unusual sliding DNA clamp and of beta-propellers in UV-damaged DNA-binding protein *Nucleic Acids Res* 28, 3570-80.

Nielsen R (1997a) A likelihood approach to populations samples of microsatellite alleles *Genetics* 146, 711-6.

Nielsen R (1997b) Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA *Syst Biol* 46, 346-53.

North JR, Morgan AJ, Thompson JL, Epstein MA (1982) Purified Epstein-Barr virus Mr 340,000 glycoprotein induces potent virus-neutralizing antibodies when incorporated in liposomes *Proc Natl Acad Sci U S A* 79, 7504-8.

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment *J Mol Biol* 302, 205-17.

Noutoshi Y, Arai R, Fujie M, Yamada T (1998) Structure of the Chlorella Zepp retrotransposon: nested Zepp clusters in the genome *Mol Gen Genet* 259, 256-63.

Novikova OS, Blinov AG (2009) [Origin, evolution, and distribution of different groups of non-LTR retrotransposons among eukaryotes] *Genetika* 45, 149-59.

Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics *Bioinformatics* 24, 581-3.

O'Brien SJ, Stanyon R (1999) Phylogenomics. Ancestral primate viewed *Nature* 402, 365-6.

Oda Y, Yoshida M, Kanaya S (1993) Role of histidine 124 in the catalytic function of ribonuclease HI from *Escherichia coli J Biol Chem* 268, 88-92.
Ohta T (1993) An examination of the generation-time effect on molecular evolution *Proc Natl Acad Sci U S A* 90, 10676-80.

Okada N, Hamada M, Ogiwara I, Ohshima K (1997) SINEs and LINEs share common 3' sequences: a review *Gene* 205, 229-43.

Oliveira R, Zani C, Ferreira R, Leite R, Alves T (2008) Síntese, avaliação biológica e modelagem molecular de arilfuranos como inibidores da enzima tripanotiona redutase *Quim. Nova* 31, 261-7.

Ondarza RN (2007) Drug targets from human pathogenic amoebas: *Entamoeba histolytica*, *Acanthamoeba polyphaga* and *Naegleria fowleri Infect Disord Drug Targets* 7, 266-80.

Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite *Nature* 284, 604-7.

Ouellette M, Legare D, Papadopoulou B (2001) Multidrug resistance and ABC transporters in parasitic protozoa *J Mol Microbiol Biotechnol* 3, 201-6.

Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers *Comput Appl Biosci* 12, 357-8.

Page RD (2007) TBMap: a taxonomic perspective on the phylogenetic database TreeBASE BMC *Bioinformatics* 8, 158.

Page RD (2002) Visualizing phylogenetic trees using TreeView *Curr Protoc Bioinformatics* Chapter 6, Unit 6 2.

Palatnik-de-Sousa CB (2008) Vaccines for leishmaniasis in the fore coming 25 years *Vaccine* 26, 1709-24.

Park J, Karplus K, Barrett C *et al*. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods *J Mol Biol* 284, 1201-10.

Pavlovic-Djuranovic S, Schultz JE, Beitz E (2003) A single aquaporin gene encodes a water/glycerol/urea facilitator in *Toxoplasma gondii* with similarity to plant tonoplast intrinsic proteins *FEBS Lett* 555, 500-4.

Peacock CS, Seeger K, Harris D *et al*. (2007) Comparative genomic analysis of three *Leishmania* species that cause diverse human disease *Nat Genet* 39, 839-47.

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA *Methods Enzymol* 183, 63-98.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles *Proc Natl Acad Sci U S A* 96, 4285-8.

Peres-Neto PR, Marques F (2000) When are random data not random, or is the PTP test useful? *Cladistics* 16, 420-4.

Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments *Mol Biol Evol* 21, 1740-52.

Pich U, Schubert I (1998) Terminal heterochromatin and alternative telomeric sequences in Allium cepa *Chromosome Res* 6, 315-21.

Plewniak F, Bianchetti L, Brelivet Y *et al*. (2003) PipeAlign: A new toolkit for protein family analysis *Nucleic Acids Res* 31, 3829-32.

Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies *Bioinformatics* 21, 676-9.

Prak ET, Kazazian HH, Jr. (2000) Mobile elements and the human genome *Nat Rev Genet* 1, 134-44.

Qian B, Goldstein RA (2004) Performance of an iterated T-HMM for homology detection *Bioinformatics* 20, 2175-80.

Quesneville H, Bergman CM, Andrieu O *et al*. (2005) Combined evidence annotation of transposable elements in genome sequences *PLoS Comput Biol* 1, 166-75.

Rannala B, Huelsenbeck JP, Yang Z, Nielsen R (1998) Taxon sampling and the accuracy of large phylogenies *Syst Biol* 47, 702-10.

Richardson JL, Nett IR, Jones DC, Abdille MH, Gilbert IH, Fairlamb AH (2009) Improved tricyclic inhibitors of trypanothione reductase by screening and chemical synthesis *ChemMedChem* 4, 1333-40.

Robbertse B, Reeves JB, Schoch CL, Spatafora JW (2006) A phylogenomic analysis of the Ascomycota *Fungal Genet Biol* 43, 715-25.

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies *Nature* 425, 798-804.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models *Bioinformatics* 19, 1572-4.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees *Mol Biol Evol* 4, 406-25.

Sanchez-Gracia A, Maside X, Charlesworth B (2005) High rate of horizontal transfer of transposable elements in *Drosophila Trends Genet* 21, 200-3.

SanMiguel P, Tikhonov A, Jin YK *et al*. (1996) Nested retrotransposons in the intergenic regions of the maize genome *Science* 274, 765-8.

Santos AL, Branquinha MH, D'Avila-Levy CM (2006) The ubiquitous gp63-like metalloprotease from lower trypanosomatids: in the search for a function *An Acad Bras Cienc* 78, 687-714.

Sauvage V, Aubert D, Escotte-Binet S, Villena I (2009) The role of ATP-binding cassette (ABC) proteins in protozoan parasites *Mol Biochem Parasitol* 167, 81-94.

Savage JM (1977) Evolution. In MB Series. ed. New York, N. Y.

Sawyer SL, Malik HS (2006) Positive selection of yeast nonhomologous end-joining genes and a retrotransposon conflict hypothesis *Proc Natl Acad Sci U S A* 103, 17614-9.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing *Bioinformatics* 18, 502-4.

Schmidt HA, von Haeseler A (2007) Maximum-likelihood analysis using TREE-PUZZLE *Curr Protoc Bioinformatics* Chapter 6, Unit 6

Schmidt A, Krauth-Siegel RL (2002) Enzymes of the trypanothione metabolism as targets for antitrypanosomal drug development *Curr Top Med Chem* 2, 1239-59.
Schneider H (2003) *Métodos de análise filogenética: um guia prático* Ribeirão Preto.

Scorzetti G, Fell JW, Fonseca A, Statzell-Tallman A (2002) Systematics of basidiomycetous yeasts: a comparison of large subunit D1/D2 and internal transcribed spacer rDNA regions *FEMS Yeast Res* 2, 495-517.

Sicheritz-Ponten T, Andersson SG (2001) A phylogenomic approach to microbial evolution *Nucleic Acids Res* 29, 545-52.

Simpson AG, MacQuarrie EK, Roger AJ (2002) Eukaryotic evolution: early origin of canonical introns *Nature* 419, 270.

Singh BK, Sarkar N, Jagannadham MV, Dubey VK (2008) Modeled structure of trypanothione reductase of *Leishmania* infantum *BMB Rep* 41, 444-7.

Sjolander K (2004) Phylogenomic inference of protein molecular function: advances and challenges *Bioinformatics* 20, 170-9.

Smit AF, Riggs AD (1996) Tiggers and DNA transposon fossils in the human genome *Proc Natl Acad Sci U S A* 93, 1443-8.

Snell EA, Furlong RF, Holland PW (2001) Hsp70 sequences indicate that choanoflagellates are closely related to animals *Curr Biol* 11, 967-70.

Sobel JD, Nagappan V, Nyirjesy P (1999) Metronidazole-resistant vaginal trichomoniasis--an emerging problem *N Engl J Med* 341, 292-3.
Sogin ML (1991) Early evolution and the origin of eukaryotes *Curr Opin Genet Dev* 1, 457-63.

Sogin ML, Gunderson JH, Elwood HJ, Alonso RA, Peattie DA (1989) Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia Science* 243, 75-7.

Spitzer M (2006) Automating the analysis of protein family evolution. (PhD-Thesis.): University of Muenster.

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees *Bioinformatics* 21, 456-63.

Stechmann A, Cavalier-Smith T (2003a) Phylogenetic analysis of eukaryotes using heat-shock protein Hsp90 *J Mol Evol* 57, 408-19.

Stechmann A, Cavalier-Smith T (2003b) The root of the eukaryote tree pinpointed *Curr Biol* 13, R665-6.

Stolle K, Schnoor M, Fuellen G *et al*. (2005) Cloning, cellular localization, genomic organization, and tissue-specific expression of the TGFbeta1-inducible SMAP-5 gene *Gene* 351, 119-30.
Stoltzfus A (1994) Origin of introns--early or late *Nature* 369, 526-7; author reply 7-8.

Sundar S, More DK, Singh MK *et al*. (2000) Failure of pentavalent antimony in visceral leishmaniasis in India: report from the center of the Indian epidemic *Clin Infect Dis* 31, 1104-7.

Suzuki Y, Gojobori T (2001a) Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b *Gene* 276, 83-7.

Suzuki Y, Nei M (2001b) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites *Mol Biol Evol* 18, 2179-85.

Swofford DL (2002) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*.

Takahashi H, Okazaki S, Fujiwara H (1997) A new family of site-specific retrotransposons, SART1, is inserted into telomeric repeats of the silkworm, *Bombyx mori Nucleic Acids Res* 25, 1578-84.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments *Syst Biol* 56, 564-77.

Tamayo EM, Iturbe A, Hernandez E *et al*. (2005) Trypanothione reductase from the human parasite *Entamoeba histolytica*: a new drug target *Biotechnol Appl Biochem* 41, 105-15.

Tarraga J, Medina I, Arbiza L *et al*. (2007) Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics *Nucleic Acids Res* 35, W38-42.

Taylor WR (1988) A flexible method to align large numbers of biological sequences *J Mol Evol* 28, 161-9.

Tehler A, Little DP, Farris JS (2003) The full-length phylogenetic tree from 1551 ribosomal sequences of chitinous fungi, Fungi *Mycol Res* 107, 901-16.

Teichmann SA, Mitchison G (1999) Is there a phylogenetic signal in prokaryote proteins? *J Mol Evol* 49, 98-107.

Telesnitsky A, Goff S (1997) Reverse transcriptase and the generation of retroviral DNA. In ea H. Varmus ed. *Retroviruses*; pp. 121-60. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Telford MJ (2007) Phylogenomics *Curr Biol* 17, R945-6.

Temin HM (1985) Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts *Mol Biol Evol* 2, 455-68.

Temin HM (1995) Genetics of retroviruses *Ann N Y Acad Sci* 758, 161-5.

Thomarat F, Vivares CP, Gouy M (2004) Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes *J Mol Evol* 59, 780-91.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice *Nucleic Acids Res* 22, 4673-80.

Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX *Curr Protoc Bioinformatics* Chapter 2, Unit 2 3.

Thornton JW, DeSalle R (2000) Gene family evolution and homology: genomics meets phylogenetics *Annu Rev Genomics Hum Genet* 1, 41-73.

Thorsness MK, White KH, Thorsness PE (2002) Migration of mtDNA into the nucleus *Methods Mol Biol* 197, 177-86.

Turmel M, Otis C, Lemieux C (2002) The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants *Mol Biol Evol* 19, 24-38.

Uzcategui NL, Szallies A, Pavlovic-Djuranovic S *et al*. (2004) Cloning, heterologous expression, and characterization of three aquaglyceroporins from *Trypanosoma brucei J Biol Chem* 279, 42669-76.

Van Walle I, Lasters I, Wyns L (2004) Align-m--a new algorithm for multiple alignment of highly divergent sequences *Bioinformatics* 20, 1428-35.

Vazquez M, Ben-Dov C, Lorenzi H, Moore T, Schijman A, Levin MJ (2000) The short interspersed repetitive element of *Trypanosoma cruzi*, SIRE, is part of VIPER, an unusual retroelement related to long terminal repeat retrotransposons *Proc Natl Acad Sci U S A* 97, 2128-33.

Vazquez MP, Schijman AG, Levin MJ (1994) A short interspersed repetitive element provides a new 3' acceptor site for trans-splicing in certain ribosomal P2 beta protein genes of *Trypanosoma cruzi Mol Biochem Parasitol* 64, 327-36.

Venter JC, Adams MD, Myers EW *et al*. (2001) The sequence of the human genome *Science* 291, 1304-51.

Vickerman K (1976) The diversity of the kinetoplastid flagellates. In ED Lumsden WH ed. *Biology of the Kinetoplastida*. London: Academic Press Inc.

Vienne A, Rasmussen J, Abi-Rached L, Pontarotti P, Gilles A (2003) Systematic phylogenomic evidence of en bloc duplication of the ancestral 8p11.21-8p21.3-like region *Mol Biol Evol* 20, 1290-8.

Villanueva MS, Williams SP, Beard CB, Richards FF, Aksoy S (1991) A new member of a family of site-specific retrotransposons is present in the spliced leader RNA genes of *Trypanosoma cruzi Mol Cell Biol* 11, 6139-48.

Wainright PO, Hinkle G, Sogin ML, Stickel SK (1993) Monophyletic origins of the metazoa: an evolutionary link with fungi *Science* 260, 340-2.

Wallace IM, O'Sullivan O, Higgins DG (2005) Evaluation of iterative alignment algorithms for multiple alignment *Bioinformatics* 21, 1408-14.

Walther TC, Kennell JC (1999) Linear mitochondrial plasmids of *F. oxysporum* are novel, telomere-like retroelements *Mol Cell* 4, 229-38.

Wang H, Lambowitz AM (1993) The Mauriceville plasmid reverse transcriptase can initiate cDNA synthesis de novo and may be related to reverse transcriptase and DNA polymerase progenitor *Cell* 75, 1071-81.

Weinrich SL, Pruzan R, Ma L *et al*. (1997) Reconstitution of human telomerase with the template RNA component hTR and the catalytic protein subunit hTRT *Nat Genet* 17, 498-502.

Whisson SC, Avrova AO, Lavrova O, Pritchard L (2005) Families of short interspersed elements in the genome of the oomycete plant pathogen, *Phytophthora infestans Fungal Genet Biol* 42, 351-65.

Wilgenbusch JC, Swofford D (2003) Inferring evolutionary trees with PAUP* *Curr Protoc Bioinformatics* Chapter 6, Unit 6 4.

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

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya *Proc Natl Acad Sci U S A* 87, 4576-9.

Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference *Genome Biol* 9, R151.

Xiao L, Fayer R (2008) Molecular characterisation of species and genotypes of *Cryptosporidium* and *Giardia* and assessment of zoonotic transmission. *Int J Parasitol* 38(11):1239-55.

Xiong Y, Eickbush TH (1988) Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns *Mol Biol Evol* 5, 675-90.

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences *Embo J* 9, 3353-62.

Xiong Y, Eickbush TH (1993) Dong, a non-long terminal repeat (non-LTR) retrotransposable element from Bombyx mori *Nucleic Acids Res* 21, 1318.

Xu P, Widmer G, Wang Y *et al*. (2004) The genome of *Cryptosporidium hominis Nature* 431, 1107-12.

Yamanaka K ST, Inouye M (). (eds. N.L. Craig, *et al*.). (2002) Retrons in Mobile DNA II. In N Craig, Craigie, R., Gellert, M., and Lambowitz, A ed. *Mobile DNA II*; pp. 784-95. Washington, DC: American Society for Microbiology (ASM Press).

Yang Z (1994a) Estimating the pattern of nucleotide substitution *J Mol Evol* 39, 105-11.

Yang Z (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods *J Mol Evol* 39, 306-14.

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites *Genetics* 155, 431-49.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood *Mol Biol Evol* 24, 1586-91.

Yardley V, Ortuno N, Llanos-Cuentas A *et al*. (2006) American tegumentary leishmaniasis: Is antimonial treatment outcome related to parasite drug susceptibility? *J Infect Dis* 194, 1168-75.

Zhu G, Keithly JS, Philippe H (2000). What is the phylogenetic position of *Cryptosporidium*? Int J Syst *Evol Microbiol* 50 Pt 4(1673-81.

Zhu Y, Zou S, Wright DA, Voytas DF (1999) Tagging chromatin with retrotransposons: target specificity of the Saccharomyces Ty5 retrotransposon changes with the chromosomal localization of Sir3p and Sir4p *Genes Dev* 13, 2738-49.

Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz AM (1995a) A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility *Cell* 83, 529-38.

Zimmerly S, Guo H, Perlman PS, Lambowitz AM (1995b) Group II intron mobility occurs by target DNA-primed reverse transcription *Cell* 82, 545-54.

Zmasek CM, Eddy SR (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree *Bioinformatics* 17, 821-8.

Zmasek CM, Eddy SR (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs *BMC Bioinformatics* 3, 14.

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Texas: The University of Texas at Austin.

CAPÍTULO 8 - ANEXOS

8.1 Termos e Conceitos

8.1.1 Filogenia

Arvore filogenética, árvore evolutiva ou filogenia representa as relações de ancestralidade comum entre táxons ou grupos. Uma filogenia sempre representa uma hipótese de relações de grupos, inferida com base nos dados estudados, os quais podem ser caracteres morfológicos ou moleculares. Esta hipótese está sujeita a ser confirmada ou refutada no futuro, assim como qualquer hipótese dentro do método científico (hipotético-dedutivo). A árvore da vida já "aconteceu" (e continua "acontecendo") e é, portanto, um evento histórico, sendo que a filogenia pode ser inferida e não aferida. A ordem em que os táxons terminais aparecem não importa. Importa quem é o ancestral comum mais recente (ACMR).

Existem dois componentes importantes que podem ser utilizados quando comparamos duas filogenias: a **topologia**, que indica a posição dos táxons na árvore, ou seja, a relação entre grupos irmãos, e os **comprimentos dos ramos**, que indicam a quantidade de mudanças que ocorrem em cada ramo da árvore.

Existem três formas básicas de representar uma filogenia: o **cladograma,** que mostra somente a topologia; o **filograma** (ou árvore aditiva), que mostra a topologia e o comprimento dos ramos; e o **dendograma** (ou árvore ultramétrica), que mostra a topologia e o comprimento dos ramos na forma de tempo absoluto.

A **árvore enraizada** possui uma raiz - o nó basal - que vai determinar a direção da transformação de caracteres. A topologia pode ser apresentada também como sendo uma **árvore não enraizada**.

Usando a terminologia de Hennig, em uma determinada árvore, um caractere pode possuir dois estados: **plesiomórfico** (ancestral ou primitivo) e **apomórfico** (derivado). Quando um estado de caractere plesiomórfico é compartilhado por dois ou mais táxons, o chamamos de **simplesiomorfia**. Quando um estado de caractere apomórfico é compartilhado por dois ou mais táxons, chamamos de **sinapomorfia**. Quando somente um dos táxons na árvore possui a condição apomórfica, chamamos de **autopomorfia**. As sinapomorfias são homologias especiais, pois agrupam todos os descendentes de um ancestral comum. O grupo formado por todos os

descendentes de um ancestral comum em uma filogenia é chamado de **monofilético** e caracterizado por sinapomorfias. Caso o grupo não inclua todos os descendentes de um ancestral, ele é chamado de **parafilético** e caracterizado por simplesiomorfias. Um grupo que reúna táxons filogeneticamente distantes com base em **homoplasias** (características que evoluíram independentemente por reversão ou paralelismo) é chamado de **polifilético**. A diferença entre grupos para e polifiléticos nem sempre é clara. Alguns autores consideram sinônimos os termos **homologia** e sinapomorfia, enquanto outros consideram tanto as sinapomorfias quanto as simplesiomorfias como homologias, pois teriam sido herdadas de um ancestral comum.

8.1.2 UTO

É a Unidade Taxonômica Operacional. Este termo foi originalmente cunhado por Sneath e Sokal {Sneath, 1962 #472; Jones, 1973 #471} (proponentes da taxonomia numérica) para denominar as unidades objeto da análise. Hoje o termo é universalmente utilizado para denominar os táxons terminais de uma árvore, podendo estes ser indivíduos, populações, espécies, gêneros ou famílias. (do inglês OTU, *Operational Taxonomic Unit*).

8.1.3 Grupo irmão

Grupos ou UTO que compartilham um ACMR.

8.1.4 Grupo interno

O grupo de estudo em uma análise filogenética. Todos os táxons incluídos para inferir as relações filogenéticas pertencem a um grupo interno.

8.1.5 Grupo externo

Teoricamente é qualquer UTO que não pertença ao grupo interno. Na prática, são táxons próximos ao grupo interno mas que não pertencem a ele e que servirão de referência na polarização dos caracteres e no enraizamento da árvore.

8.1.6 Politomia

Refere-se à ausência da dicotomia, ou seja, quando um ancestral comum dá origem a mais de dois descendentes. Isto significa que qualquer relação de parentesco é possível entre os táxons

envolvidos. As politomias podem ser: a) "macias" (*soft polytomy*), ou seja, resultado de dados insuficientes ou conflitantes e, portanto, um artefato da técnica ou b) "duras" (*hard polytomy*), que indicam um evento real de especiação praticamente simultânea de mais de dois grupos, como no caso de uma radiação adaptativa, por exemplo. Quando existem três grupos envolvidos denomina-se tricotomia.

8.1.7 Filogenia

Refere-se à história evolutiva, à história da descendência e às modificações que podem ser apresentadas dependendo do tipo de seleção. O termo também é utilizado às vezes para uma hipótese de relações filogenéticas (como em referência ao termo filogenia).

8.1.8 Árvore Filogenética

Diagrama que mostra as linhagens e as relações dos organismos. Refere-se à representação gráfica da homologia.

8.1.9 Filoma

Um grupo completo de árvores filogenéticas derivadas do proteoma de um organismo {Sicheritz-Ponten, 2001 #277}.

8.1.10 Homologia

Refere-se à relação entre sequências que compartilham uma sequência ancestral comum.

8.1.11 Filogenômica

Tem sido definida essencialmente como a interseção entre a evolução e a genômica.

8.2 Reconstrução automática de análises filogenômicas (ARPA) dos genes relacionados à resistência a drogas em genomas de protozoários

Anexo 8.1 - Lista dos seis genes candidatos a drogas
Os genes foram obtidos dos bancos de dados NCBI (COG/KOG) e PFAM

Banco de dados	AQP	hsp70	TRYR	GP63	MRPA
COG	*	*	*	*	*
KOG	KOG0223 KOG0224	KOG0101	*	*	KOG0054
PFAM	PF00230	PF00012	PF00070 PF07992 PF02852	*	PF00005 PF00664
PSI-BLAST	250 seq.	250 seq.	250 seq.	250 seq.	250 seq.
***PFAM fs hmm**	PF00230	PF00012	PF00070 PF07992 PF02852	*	PF00005 PF00664
***PFAM ls hmm**	PF00230	PF00012	PF00070 PF07992 PF02852	*	PF00005 PF00664

* Sequências não encontradas no banco de dados.
* PFAM fs hmm e ls hmm são perfis HMM (Hidden Markov Models) (Eddy, 1998) obtidos do banco de dados PFAM.

Anexo 8.2-A - Lista dos melhores *hits* do hmmpfam incluídos na filogenia
As espécies estão representadas pelas abreviaturas respectivas nas árvores filogenéticas e o número de GI

AQP		GP63		hsp70		TRYR		MRPA	
Abrev.	**GI**	**Abrev.**	**GI**	**Abrev.**	**GI**	**Abrev.**	**GI**	**Abrev.**	**GI**
YEAST	O93938	DROSO	Q9VH19	YEAST	6322426	22Lbr	134059170	124Tva	123405896
80Ddi	6012182	279Ddi	37622955	648Acu	53829570	28Lin	134067080	144Tva	123446330
60Edi	167380957	249Edi	167393821	684Bbo	6492134	48Tcr	14599976	183Pfa	124506379
88Ehi	67476549	325Ehi	67473765	817Bcab	83715961	49Tcr	14599978	190Pfa	124809620
6Lbr	134062088	281Lae	39104250	15Bcac	111120209	50Tcr	14599980	211Lbr	134062148
44Ldo	148533557	1Lam	1100213	17Bcar	111120213	51Tcr	14599982	235Lin	134069806
39Lin	146094399	280Lar	39104248	251Bcav	146134154	52Tcr	14599984	240Lin	134071361
75Lma	45645039	101Lbr	134059821	359Bdi	164458432	53Tcr	14599986	243Lin	134071789
79Lme	57792260	248Ldo	159335	24Beq	111120227	54Tcr	14599988	250Ddi	1399915
72Lta	33355657	288Ldoc	5679610	308Bgi	156511059	55Tcr	14599990	353Lin	146087038
63Mbr	167536710	247Lgu	159331	487Bmi	1842232	56Tcr	14599992	358Lin	146093772
15Pte	145477465	52Lin	1213330	21Bod	111120221	57Tcr	14599994	362Lin	146094925
89Pbe	68066408	267Lma	2761012	201Bor	145286435	58Tcr	14599996	377Tth	146164637
91Pchc	70948126	479Lme	9560	22Bov	111120223	59Tcr	14599998	395Pyo	15429020
4Pfa	124804459	265Lpa	2746321	550Bro	28465379	60Tcr	14600000	406Lbr	154337750
66Pkn	193809405	472Lsp	78125986	367Bsp	164458448	61Tcr	14600002	429Tva	154416691
50Pvi	156098853	258Ltr	21954468	351Bna	161899613	62Tcr	14600004	444Pvi	156095386
103Pyoy	82597036	260Ltu	21954472	528Can	22128362	63Tcr	14600006	450Pvi	156100371
1Tth_	118376942	54Tva	123345028	837Cba	8515216	64Tcr	14600008	466Lma	157869596
71Tgo	30575793	467Tbr	71756043	672Cbo	58012995	65Tcr	14600010	472Lma	157873201
99Tbr	71749114	261Tbrr	2196914	500Cca	193083705	66Tcr	14600012	501Gla	159119918
76Tbrb	46518239	275Tcr	31322786	155Cfe	12744806	67Tcr	14600014	503Ldo	159363
69Tcr	21717526	2Tth	118345550	549Cga	28192564	68Tcr	14600016	504Len	159370
				628Cho	38639371	69Tcr	14600018	505Ldo	159372

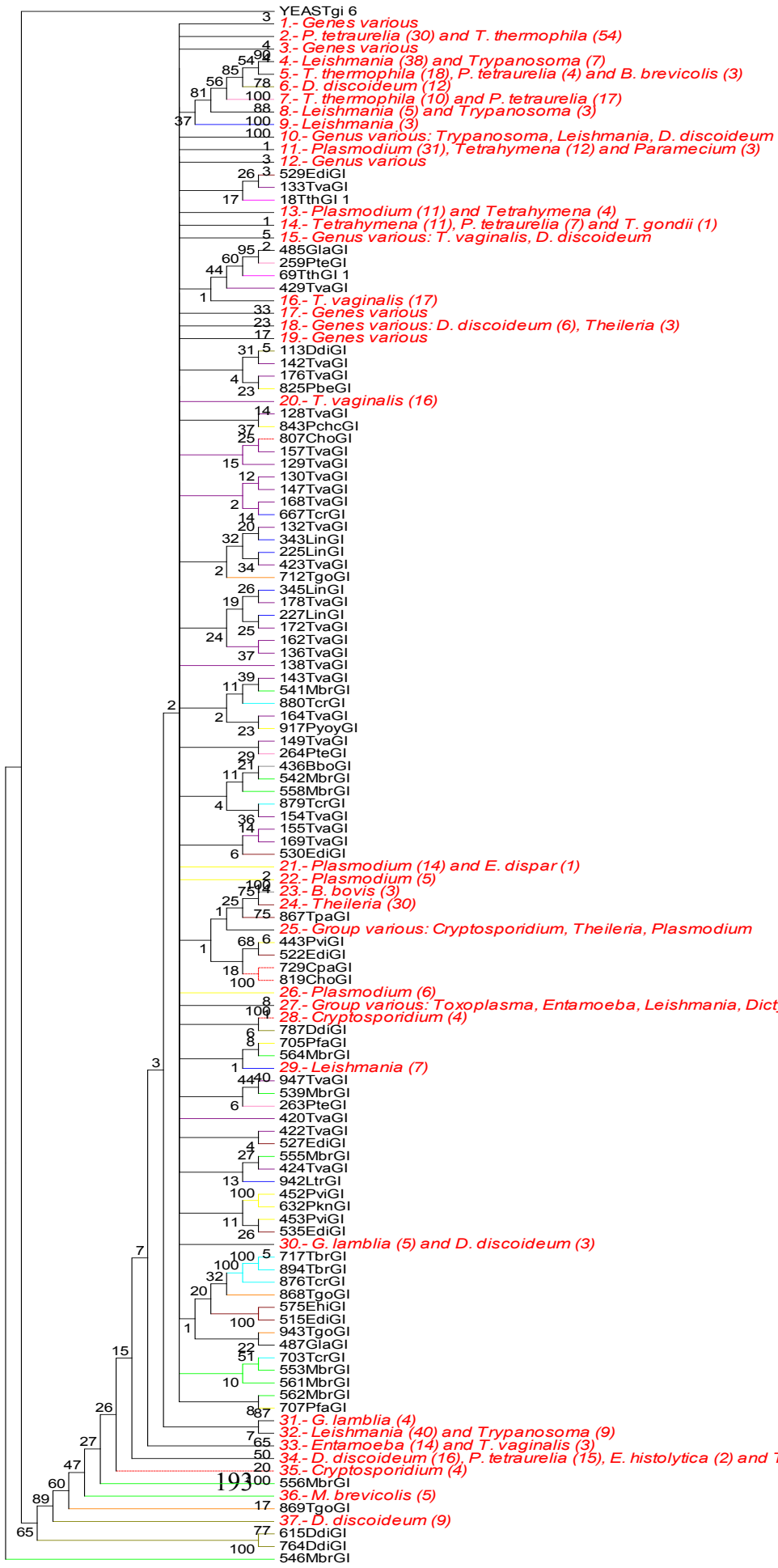
810Cme	76574006	70Tcr	14600020	510Edi	167375582
840Cmu	8515222	71Tcr	14600022	512Edi	167376553
842Cpa	88659553	72Tcr	14600024	514Edi	167377245
839Cse	8515220	73Tcr	14600026	518Edi	167382273
7Csp	10834634	74Tcr	14600028	522Edi	167385657
37Csu	115583755	75Tcr	14600030	527Edi	167387335
835Cwr	8515210	76Tcr	14600032	529Edi	167389783
562Cmer	30468181	77Tcr	14600034	530Edi	167389787
31Ccal	11465421	78Tcr	14600036	532Edi	167394252
152Cypa	126165882	79Tcr	14600038	535Edi	167395125
543Ddi	27414161	80Tcr	14600040	536Edi	167395240
630Eac	401829	81Tcr	14600042	575Ehi	183232088
673Ema	603812	82Tcr	14600044	578Ehi	183232963
674Ete	603814	83Tcr	14600046	626Pkn	193809653
783Ehu	71842313	84Tcr	14600048	630Pkn	193810467
649Edi	54397731	85Tcrm	14600050	653Pch	22002636
712Ehi	67474975	86Tcrm	14600052	658Pfa	23497526
806Ein	7579069	87Tve	14600054	659Pfa	23504716
200Gin	13569206	88Lin	146076772	660Ltr	2360941
332Gla	159114951	96Lam	148283954	666Pfa	294167
646Gtel	51209944	98Lbr	154332065	669Pfa	31540682
34Gth	11467716	114Lma	157863924	670Pfa	31540684
341Han	160331065	122Ldod	16076075	671Lma	32330114
489Hak	189095435	123Tco	162311	678Pfa	4261570
554Lam	293057	124Tcr	162317	681Pfa	439854
609Lbr	3386646	150Ehi	20799665	682Lam	4521245
485Lch	184132914	161Ldo	312821	685Lam	46401536
647Ldo	51847751	168Tcr	552297	687Pvi	47834377
807Lin	758136	169Tcr	61207340	692Pvi	51511430
336Lma	159356	170Tcr	61207342	693Pvi	51511432
622Lta	37813097	171Tcr	61207344	694Pvi	51511434
819Mja	84105385	172Tcr	61207346	695Pvi	51511436
432Mbr	167527877	173Tcr	61207348	696Pvi	51511438
73Mov	12007351	174Tcr	61207350	697Pvi	51511440
645Nfo	51012459	175Tcr	61207352	698Pvi	51511442
608Ngr	33694258	176Tcr	61207354	699Pvi	51511444
33Osi	11467472	177Tcr	61207356	700Pvi	51511446
591Pbu	33235656	178Tcr	61207358	701Pvi	51511448
162Pca	13359317	180Tcr	61207360	706Tpy	55467210
598Pdu	33235670	181Tcr	61207362	707Pfa	55925067
599Pje	33235672	182Tcr	61207364	708Pfa	55925077
600Pmu	33235674	183Tcr	61207366	709Pfa	55925091
597Pne	33235668	184Tcr	61207368	710Pfa	55925099
589Pno	33235652	185Tcr	61207370	711Pfa	55925107
588Poc	33235650	186Tcr	61207372	718Tbr	62358987
586Ppe	33235646	187Tcr	61207374	721Tbr	62360100
595Ppo	33235664	188Tcr	61207376	825Pbe	68066210
584Ppr	33235642	189Tcr	61207378	829Pbe	68076009
593Ppu	33235660	193Tcr	624038	847Pchc	70942116
32Ppur	11465781	222Tcr	71397103	880Tcr	71406202
587Pse	33235648	223Tcr	71400296	894Tcr	71652999
570Pte	3169849	233Tbr	71748410	898Tbr	72392497
590Ptr	33235654	256Lam	104745490	905Pfa	8050813
60Ptri	118411057	257Tbr	10545	916Pyoy	83314634
678Pbe	62083585			942Ltr	85542862
650Pch	552198				
739Pchc	70924562				
343Pcy	160350				
147Pfa	124506906				
504Pkn	193809418				
808Pvi	76162902				
812Pyoy	82594363				
847Pye	90994506				
820Ram	84105387				
261Rsa	149072046				
516Tth	19879267				
61Tps	118411189				
349Tan	161869				
29Tce	111120237				
811Tle	77455531				
28Tov	111120235				
745Tpa	71029352				
638Tse	436763				
671Tgo	5738968				
111Tva	123468246				
559Tbr	295365				
539Tco	25553516				
357Tcr	162158				
63Tra	119394467				

Anexo 8.2-B - Lista das espécies de protozoários incluídas na filogenia

As espécies estão representadas pelas abreviaturas e a cor de ramos respectivos nas árvores filogenéticas

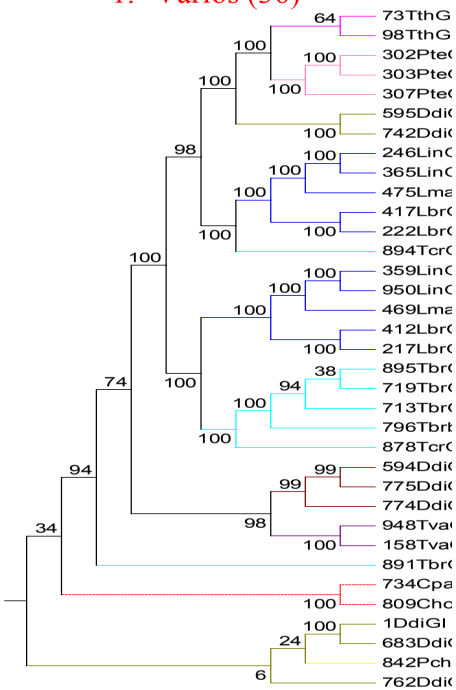
Espécie	Abrev.	Espécie	Abrev.	Espécie	Abrev.	Espécie	Abrev.	Espécie	Abrev.
Acanthamoeba culbertsoni	Acu	*Cryptosporidium parvum*	Cpa	*Leishmania braziliensis*	Lbr	*Paramecium multimicronucleatum*	Pmu	*Rhodomonas salina*	Rsa
Babesia bovis	Bbo	*Cryptosporidium serpentis*	Cse	*Leishmania chagasi*	Lch	*Paramecium nephridiatum*	Pne	*Tetrahymena pyriformis*	Tpy
Babesia caballi	Bcab	*Cryptosporidium sp*	Csp	*Leishmania donovani*	Ldo	*Paramecium novaurelia*	Pno	*Tetrahymena thermophila*	Tth
Babesia canis canis	Bcac	*Cryptosporidium suis*	Csu	*Leishmania donovani*	Ldoc	*Paramecium octaurelia*	Poc	*Thalassiosira pseudonana*	Tps
Babesia canis rossi	Bcar	*Cryptosporidium wrairi*	Cwr	*Leishmania donovani*	Ldod	*Paramecium pentaurelia*	Ppe	*Theileria annulata*	Tan
Babesia canis vogeli	Bcav	*Cyanidioschyzon merolae*	Cmer	*Leishmania enriettii*	Len	*Paramecium polycaryum*	Ppo	*Theileria cervi*	Tce
Babesia divergens	Bdi	*Cyanidium caldarium*	Ccal	*Leishmania guyanensis*	Lgu	*Paramecium primaurelia*	Ppr	*Theileria lestoquardi*	Tle
Babesia equi	Beq	*Cyanophora paradoxa*	Cypa	*Leishmania infantum*	Lin	*Paramecium putrinum*	Ppu	*Theileria ovis*	Tov
Babesia gibsoni	Bgi	*Dictyostelium discoideum*	Ddi	*Leishmania major*	Lma	*Paramecium sexaurelia*	Pse	*Theileria parva*	Tpa
Babesia microti	Bmi	*Eimeria acervulina*	Eac	*Leishmania mexicana*	Lme	*Paramecium tetraurelia*	Pte	*Theileria sergenti*	Tse
Babesia odocoilei	Bod	*Eimeria maxima*	Ema	*Leishmania panamensis*	Lpa	*Paramecium tredecaurelia*	Ptr	*Toxoplasma gondii*	Tgo
Babesia orientalis	Bor	*Eimeria tenella*	Ete	*Leishmania sp*	Lsp	*Phaeodactylum tricornutum*	Ptri	*Trichomonas vaginalis*	Tva
Babesia ovis	Bov	*Emiliania huxleyi*	Ehu	*Leishmania tarentolae*	Lta	*Physarum polycephalum*	Ppol	*Trypanosoma brucei*	Tbr
Babesia rodhaini	Bro	*Entamoeba dispar*	Edi	*Leishmania tropica*	Ltr	*Plasmodium berghei*	Pbe	*Trypanosoma brucei brucei*	Tbrb
Babesia sp	Bsp	*Entamoeba histolytica*	Ehi	*Leishmania turanica*	Ltu	*Plasmodium chabaudi*	Pch	*Trypanosoma brucei*	Tbrr
Bigelowiella natans	Bna	*Entamoeba invadens*	Ein	*Malawimonas jakobiformis*	Mja	*Plasmodium chabaudi chabaudi*	Pchc	*Trypanosoma congolense*	Tco
Cryptosporidium andersoni	Can	*Giardia intestinalis*	Gin	*Monosiga brevicolis*	Mbr	*Plasmodium cynomolgi*	Pcy	*Trypanosoma cruzi*	Tcr
Cryptosporidium baileyi	Cba	*Giardia lamblia*	Gla	*Monosiga ovata*	Mov	*Plasmodium falciparum*	Pfa	*Trypanosoma cruzi*	Tcrm
Cryptosporidium bovis	Cbo	*Gracilaria tenuistipitata*	Gtel	*Naegleria fowleri*	Nfo	*Plasmodium knowlesi*	Pkn	*Trypanosoma rangeli*	Tra
Cryptosporidium canis	Cca	*Guillardia theta*	Gth	*Naegleria gruberi*	Ngr	*Plasmodium vivax*	Pvi	*Trypanosoma vespertilionis*	Tve
Cryptosporidium felis	Cfe	*Hemiselmis andersenii*	Han	*Odontella sinensis*	Osi	*Plasmodium yoelii*	Pyo		
Cryptosporidium galli	Cga	*Heterosigma akashiwo*	Hak	*Paramecium bursaria*	Pbu	*Plasmodium yoelii yoelii*	Pyoy		
Cryptosporidium hominis	Cho	*Leishmania aethiopica*	Lae	*Paramecium caudatum*	Pca	*Porphyra purpurea*	Ppur		
Cryptosporidium meleagridis	Cme	*Leishmania amazonensis*	Lam	*Paramecium duboscqui*	Pdu	*Porphyra yezoensis*	Pye		
Cryptosporidium muris	Cmu	*Leishmania arabica*	Lar	*Paramecium jenningsi*	Pje	*Reclinomonas americana*	Ram		

MRPA - TOTAL
PAUP - AV



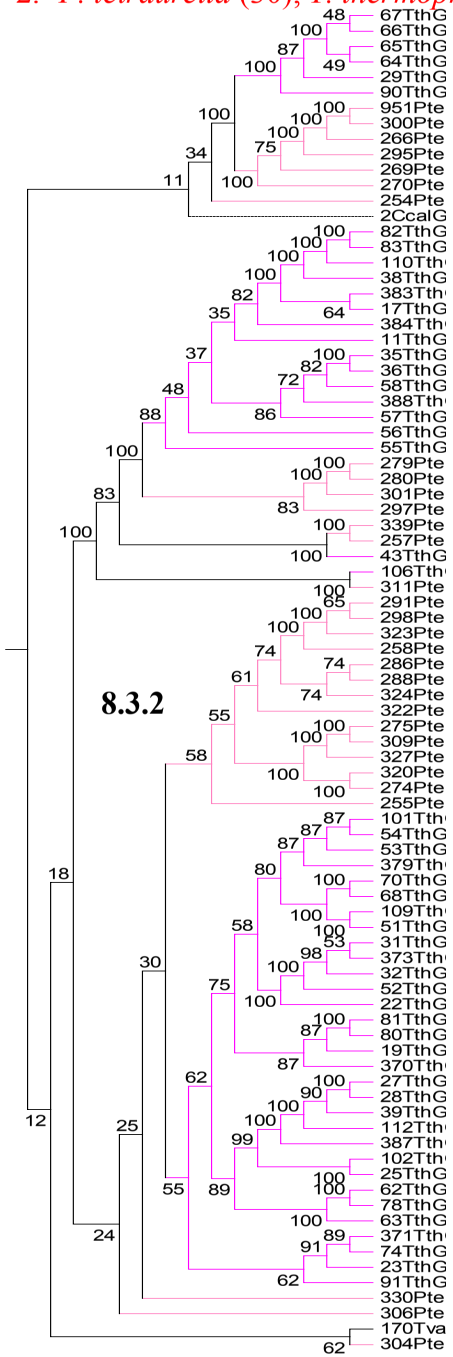
8.3

1.- Vários (36)



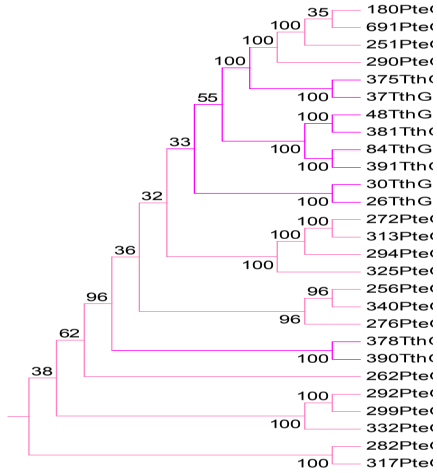
8.3.1

2.- *P. tetraurelia* (30), *T. thermophila* (54)



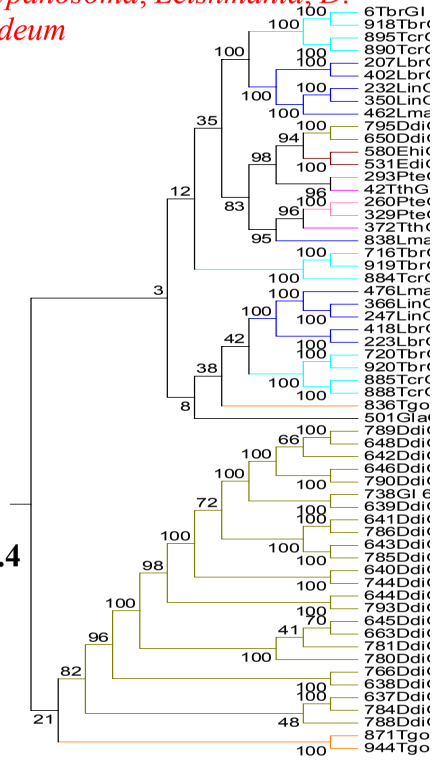
8.3.2

3.- *T. thermophila* (10), *P. tetraurelia*(17)



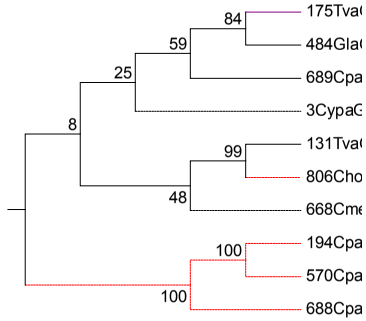
8.3.3

4.- *Trypanosoma*, *Leishmania*, *D. discoideum*



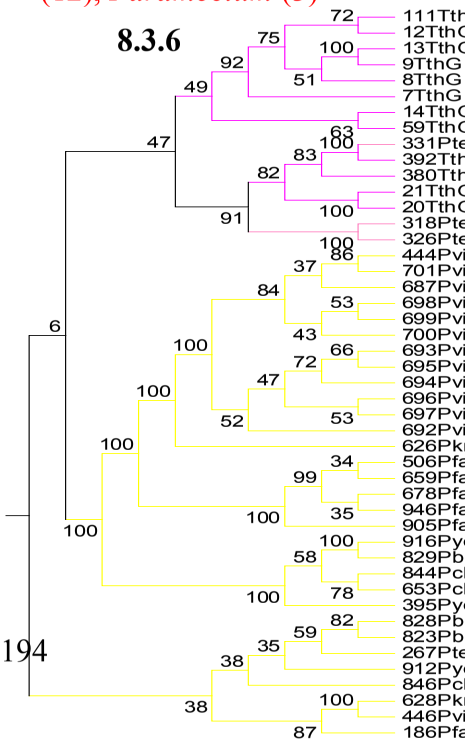
8.3.4

5.- Vários (10)



8.3.5

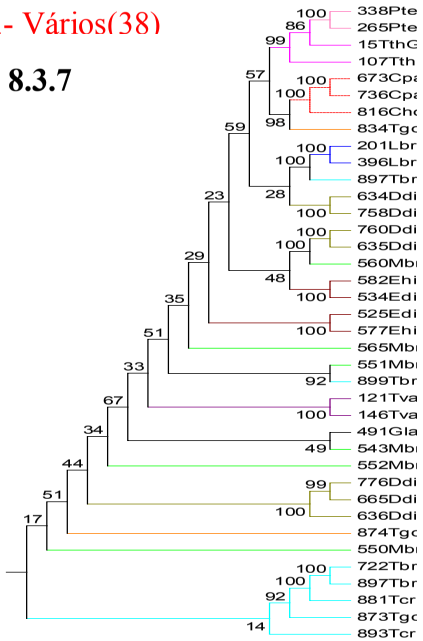
6.- *Plasmodium* (31), *Tetrahymena* (12), *Paramecium* (3)



194

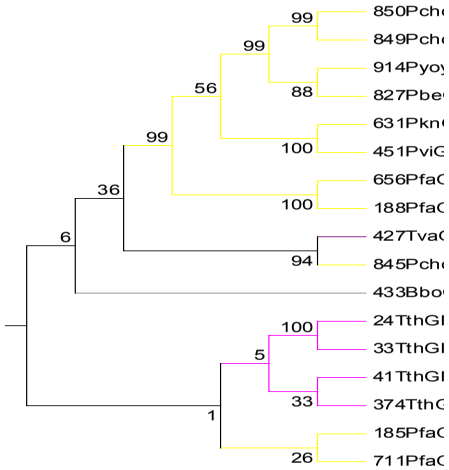
8.3.6

7.- Vários(38)



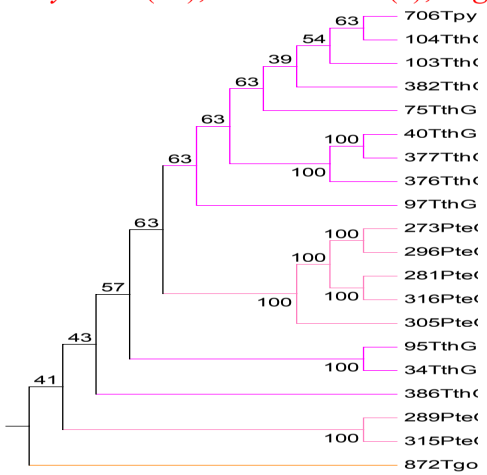
8.3.7

8.- *Plasmodium* (11), *Tetrahymena* (4)



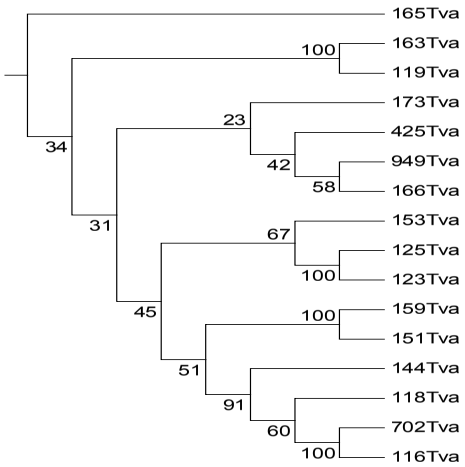
8.3.8

9.- *Tetrahymena* (11), *P. tetraurelia*(7), *T. gondii* (1)



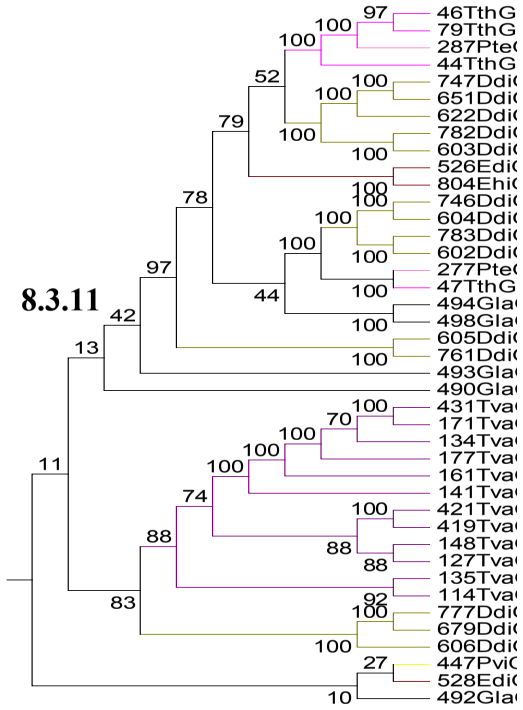
8.3.9

10.- *T. vaginalis* (17)



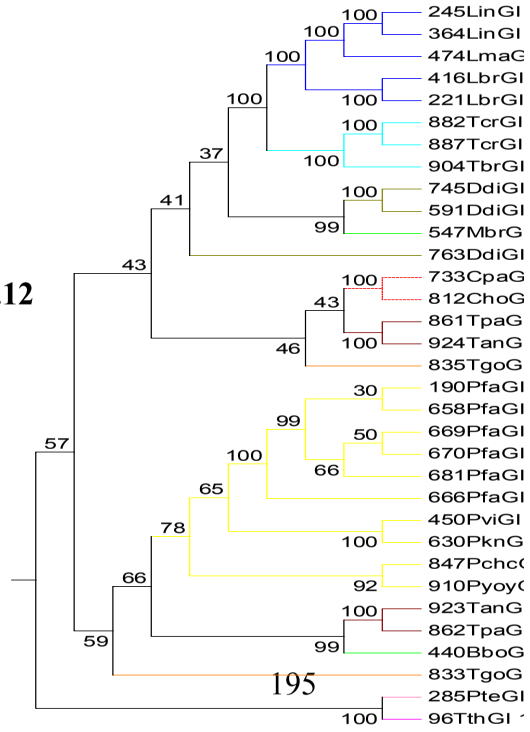
8.3.10

11.- Vários, *T. vaginalis*(12), *D. discoideum*(14)



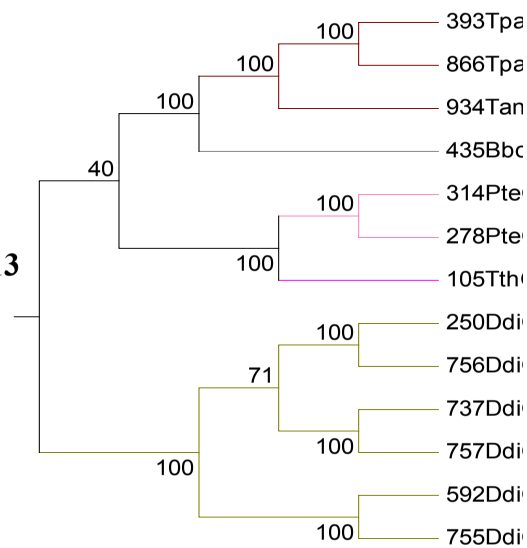
8.3.11

12.- Vários (33)



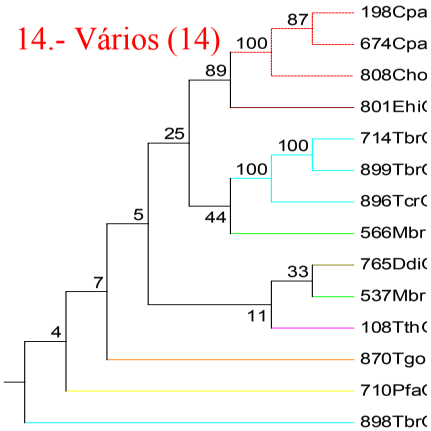
8.3.12

13.- Vários, *D. discoideum* (6), *Theileria* (3)



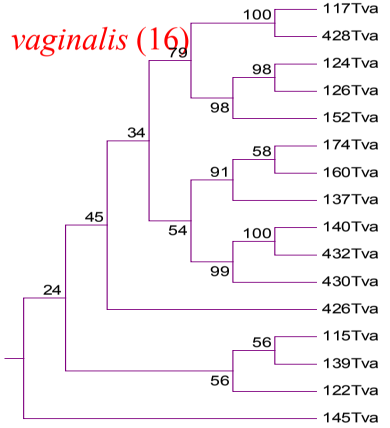
8.3.13

14.- Vários (14)



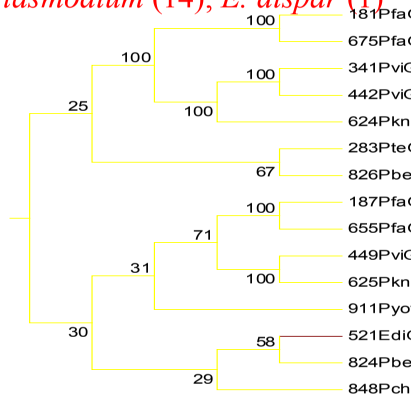
8.3.14

15.- *T. vaginalis* (16)



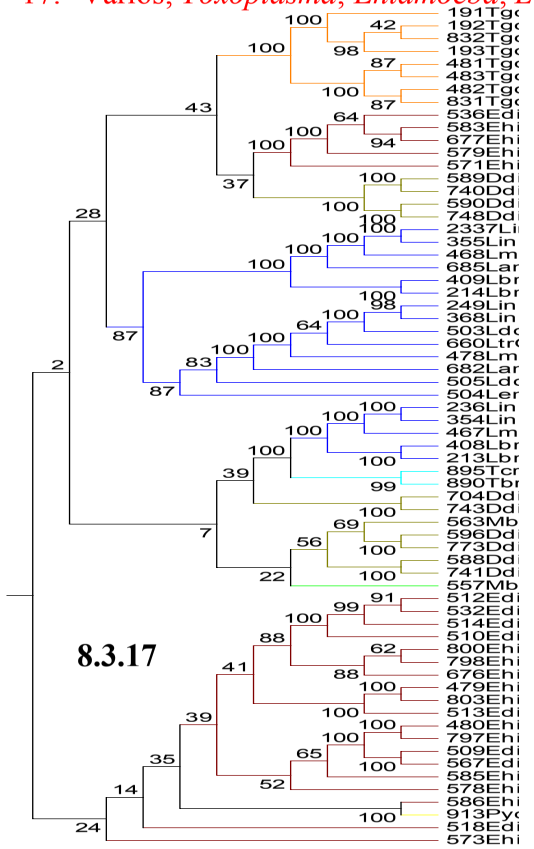
8.3.15

16.- *Plasmodium* (14), *E. dispar* (1)



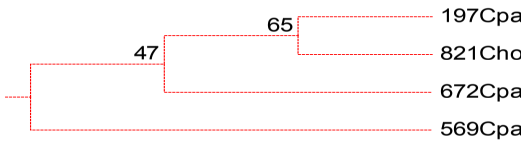
8.3.16

17.- Vários, *Toxoplasma*, *Entamoeba*, *Leishmania*, *Dictyostelium*



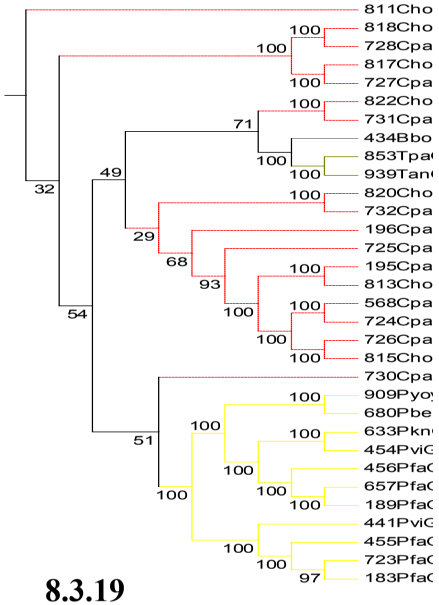
8.3.17

18.- *Cryptosporidium* (4)



8.3.18

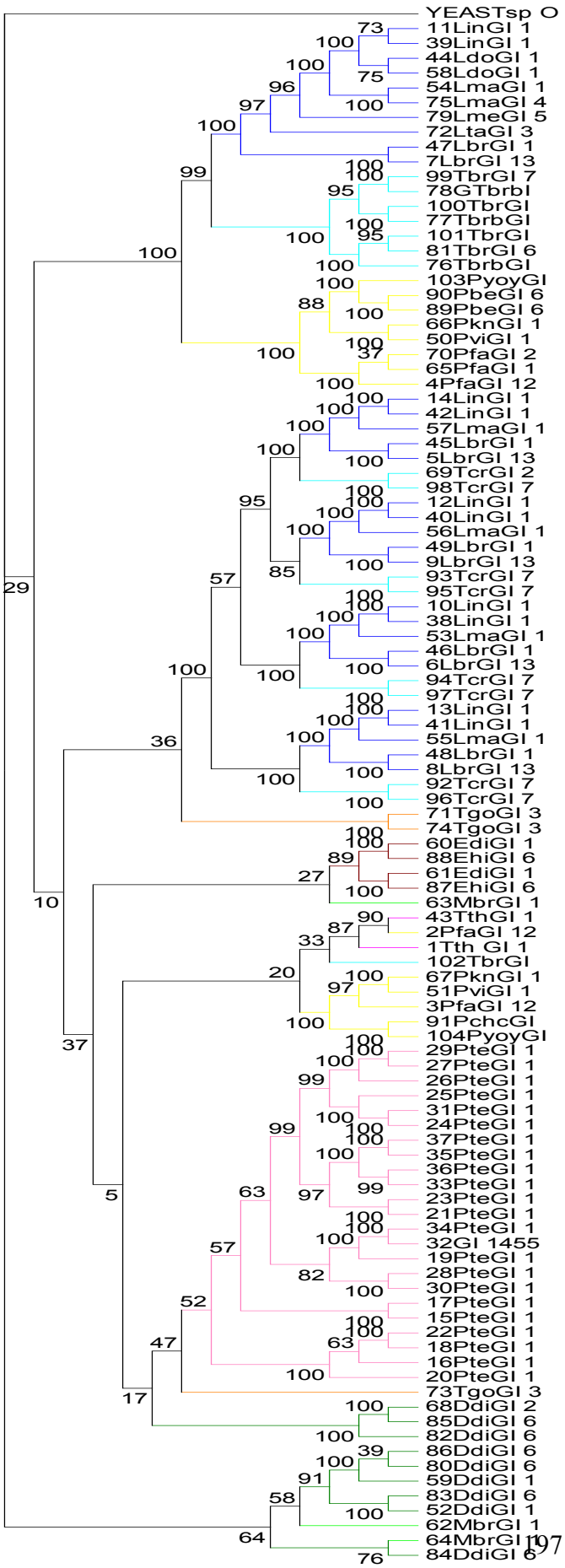
19.- Vários,*Cryptosporidium*, *Theileria*,*Plasmodium*



8.3.19

Anexo 8.3 - Árvores filogenéticas AV construídas com PAUP para o gene MRPA

As árvores filogenéticas construídas segundo a Metodologia M1 para os genes candidatos de resistência às drogas. As árvores 4.1 representam ao gene MRPA.

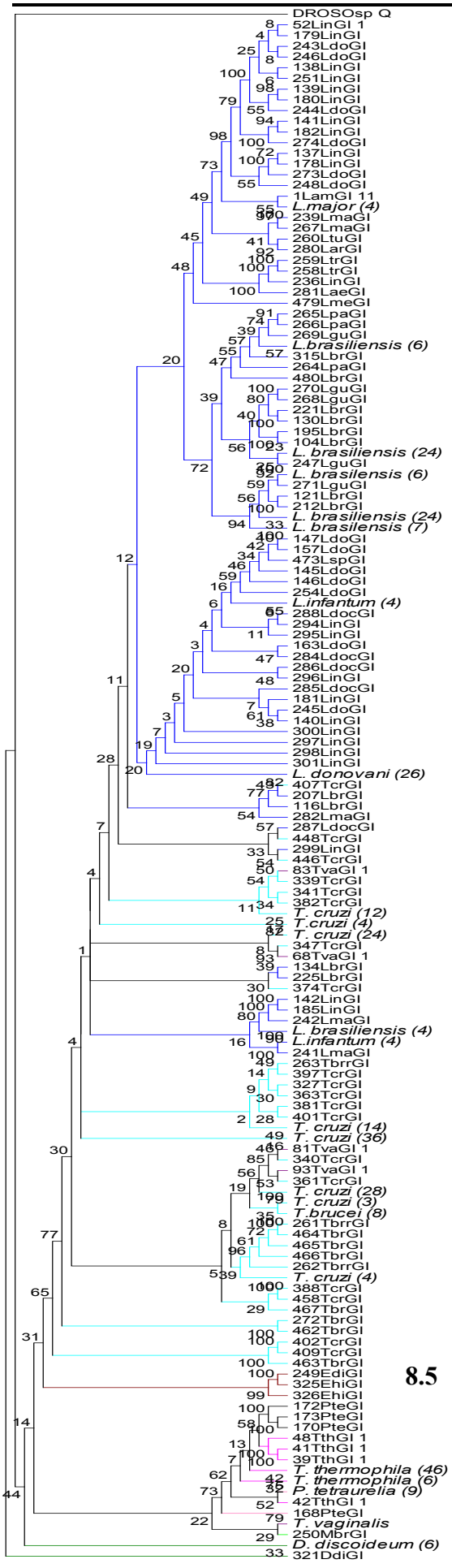


AQP - TOTAL
PAUP - AV

Anexo 8.4 - Árvores filogenéticas AV construídas com PAUP para o gene AQP

As árvores filogenéticas construídas segundo a Metodologia M1 para os genes candidatos de resistência às drogas. A árvore 4.2 representa ao gene AQP.

8.4



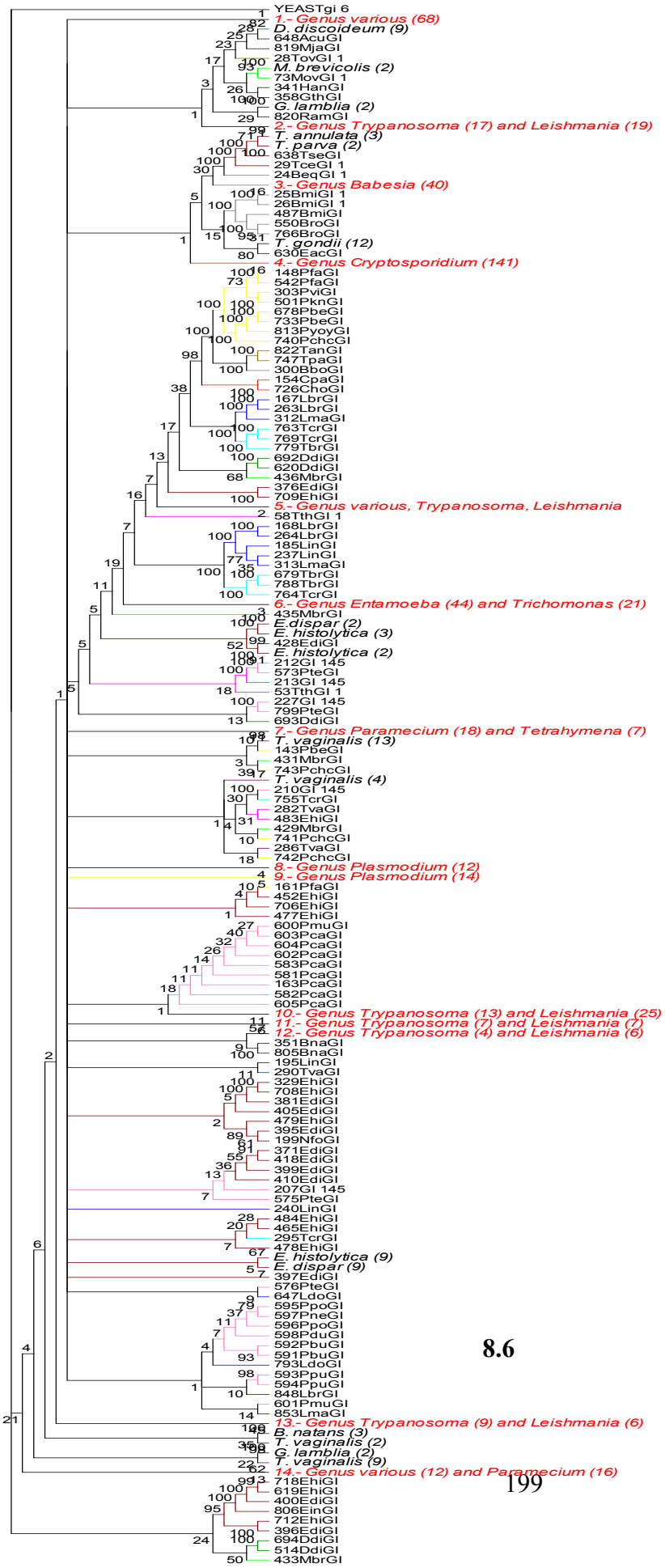
8.5

GP63 - TOTAL
PAUP - AV

Anexo 8.5 - Árvores filogenéticas AV construídas com PAUP para o gene GP63

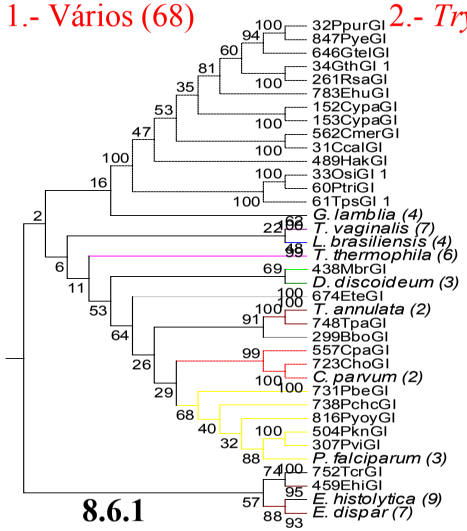
As árvores filogenéticas construídas segundo a Metodologia M1 para os genes candidatos de resistência às drogas. A árvore 4.2 representa ao gene GP63.

hsp70 - TOTAL
PAUP - AV



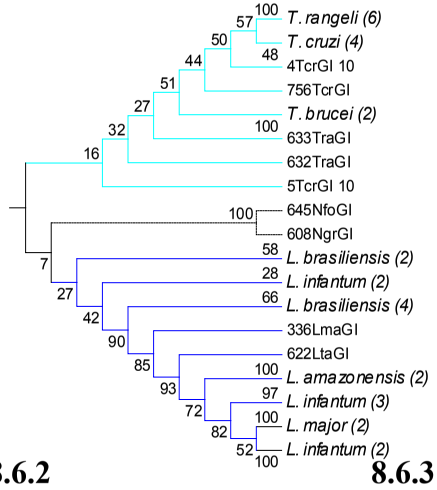
8.6

1.- Vários (68)



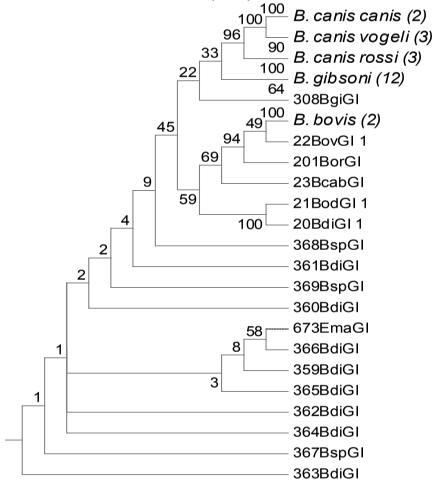
8.6.1

2.- *Trypanosoma* (17), *Leishmania* (19)



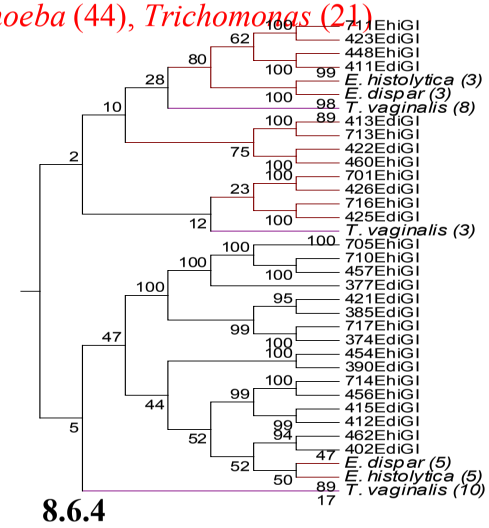
8.6.2

3.- *Babesia* (40)



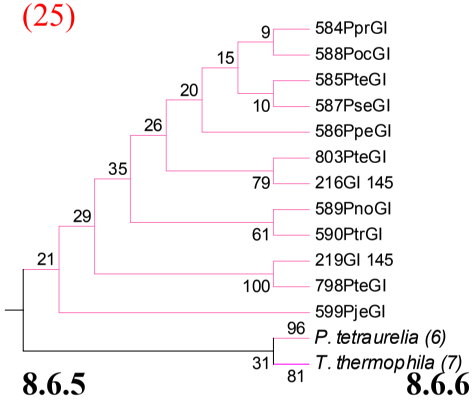
8.6.3

4.- *Entamoeba* (44), *Trichomonas* (21)



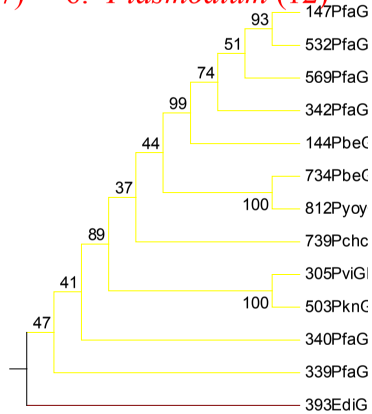
8.6.4

5.- *Paramecium* (18), *Tetrahymena* (7) (25)



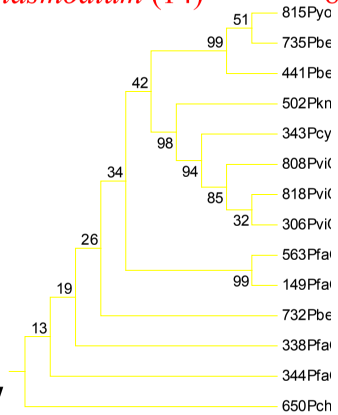
8.6.5

6.- *Plasmodium* (12)



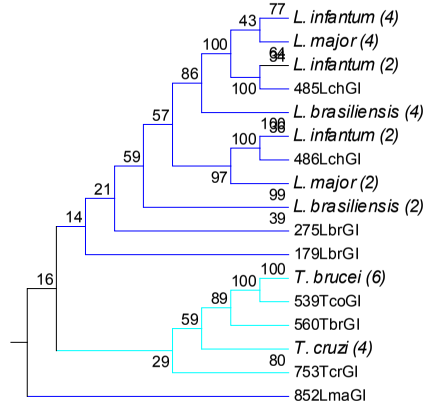
8.6.6

7.- *Plasmodium* (14)



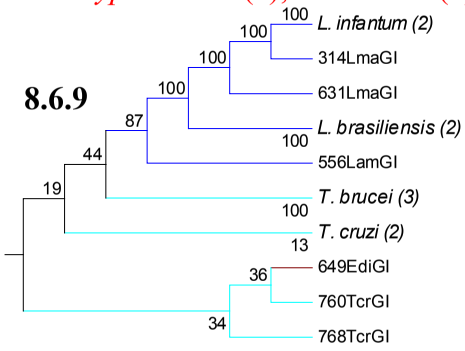
8.6.7

8.- *Trypanosoma* (13), *Leishmania*



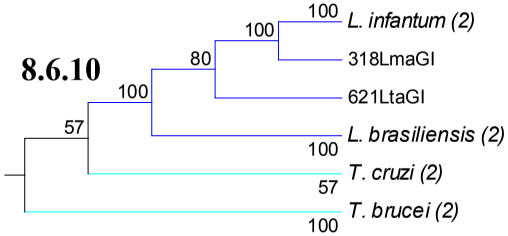
8.6.8

9.- *Trypanosoma*(7),*Leishmania*(7)



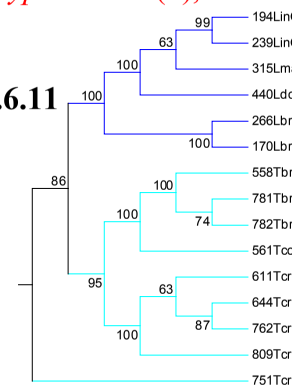
8.6.9

10.- *Trypanosoma*(4), *Leishmania*(6)



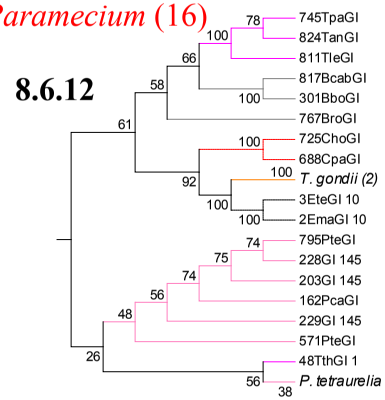
8.6.10

11.- *Trypanosoma*(9), *Leishmania*(6)



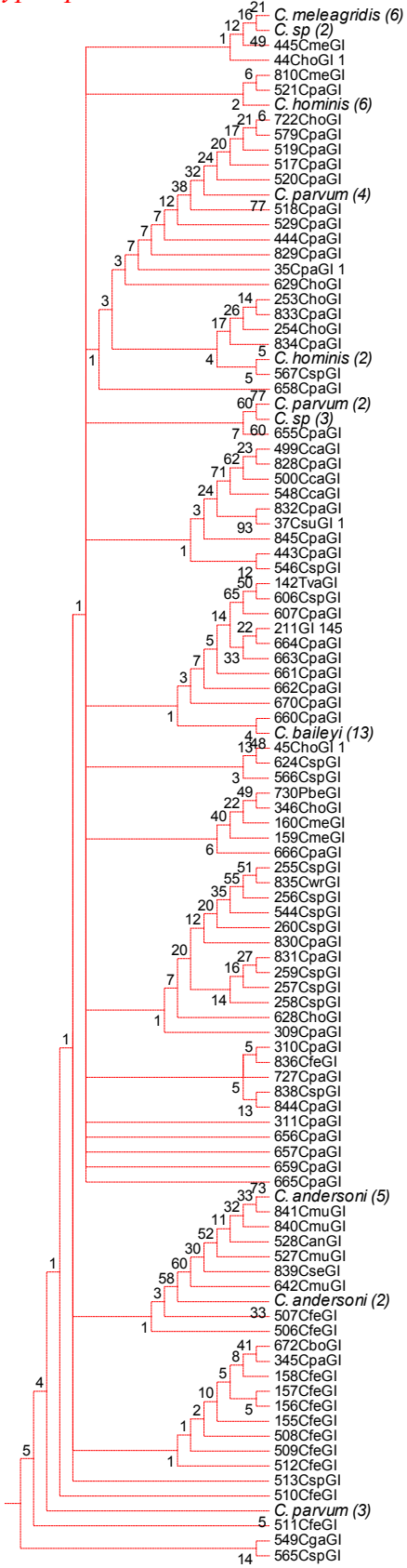
8.6.11

12.- Vários(12), *Paramecium* (16)



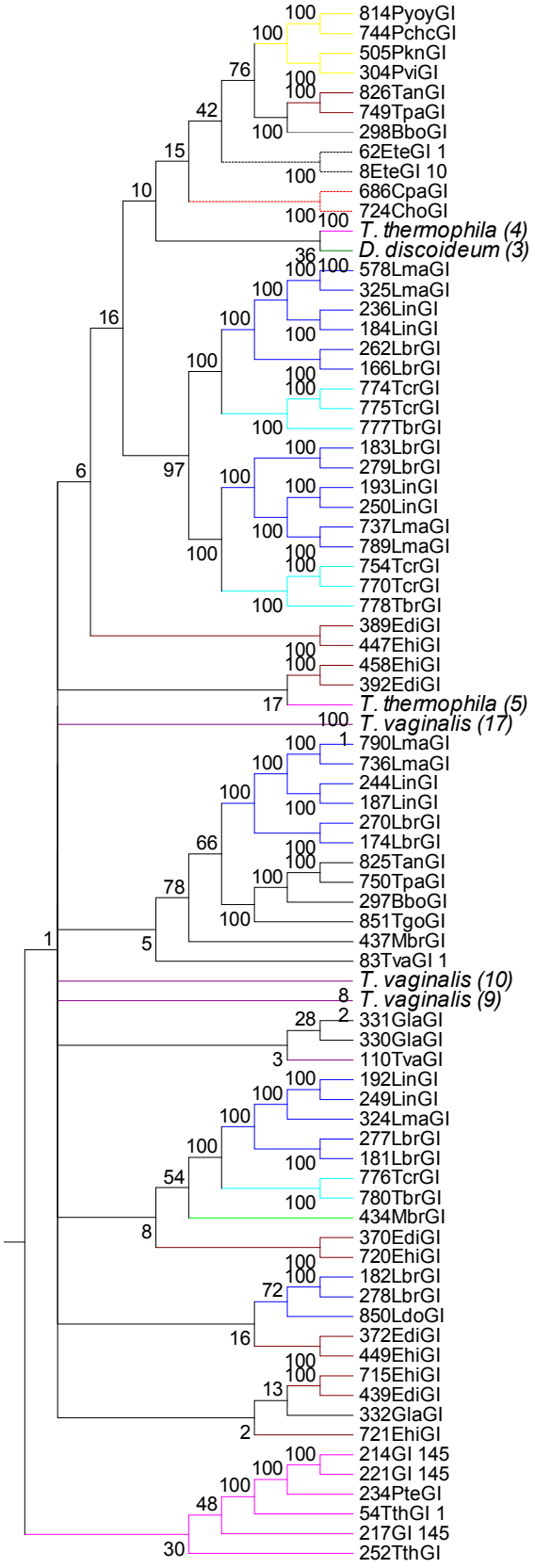
8.6.12

13.- *Cryptosporidium*



8.6.13

14.- Vários, *Trypanosoma*, *Leishmania*



8.6.14

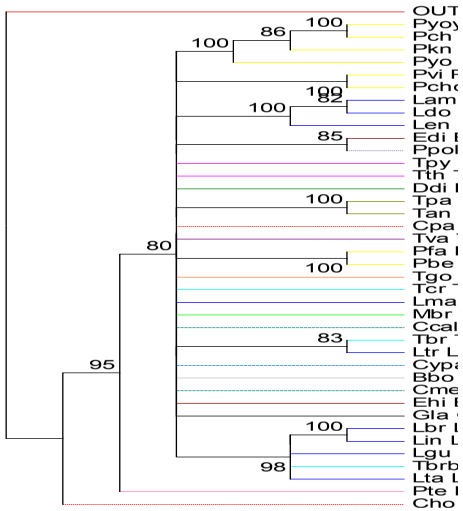
Anexo 8.6 - Árvores filogenéticas AV construídas com PAUP para o gene hsp70

As árvores filogenéticas construídas segundo a Metodologia M1 para os genes candidatos de resistência às drogas.

MRPA - Todas as Sequências Completas

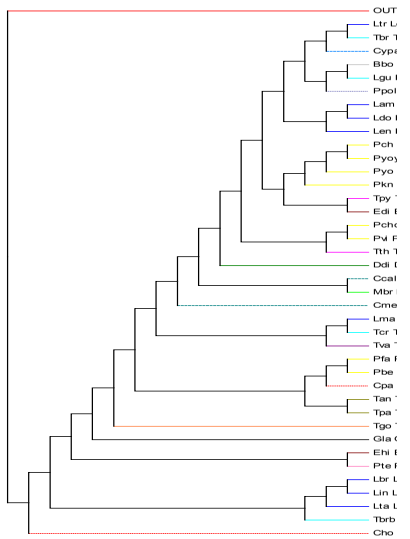
Topologia

PHYML



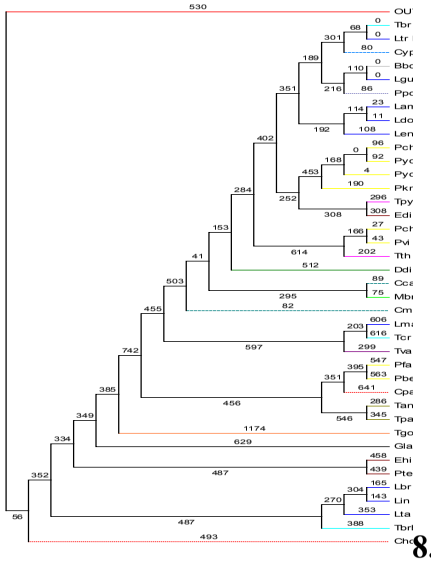
8.7.1

PAUP-AV



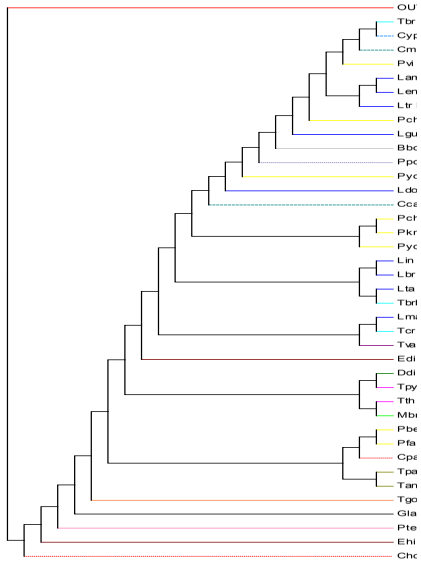
8.7.2

PAUP-MP



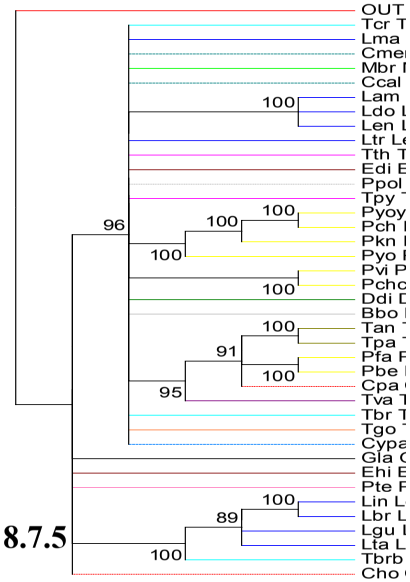
8.7.3

WEIGHBOR

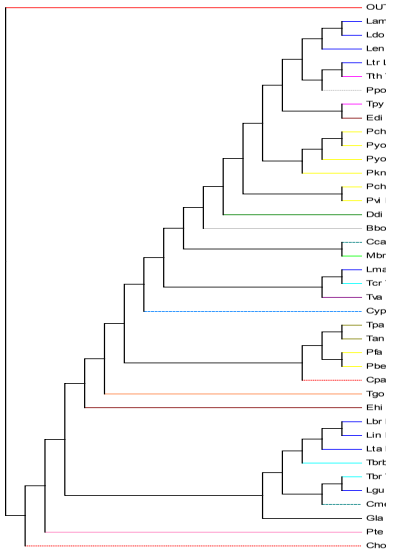


8.7.4

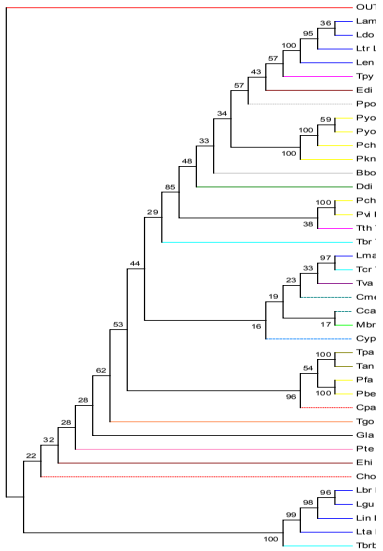
MRPA - Todas as Sequências Trimadas com o TRIMAL



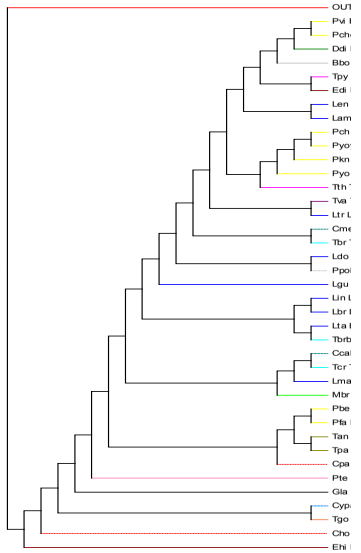
8.7.5



8.7.6



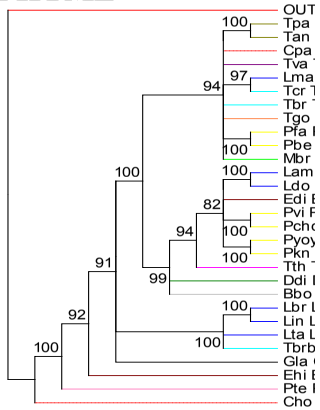
8.7.7



8.7.8

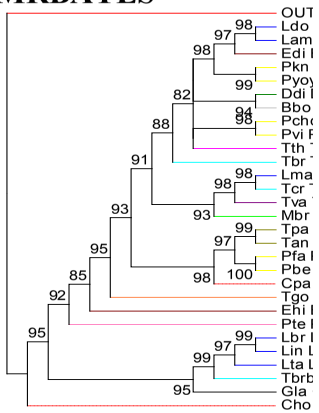
MRPA - Quatro Sequências Completas
Mostrar só topologia

PHYML



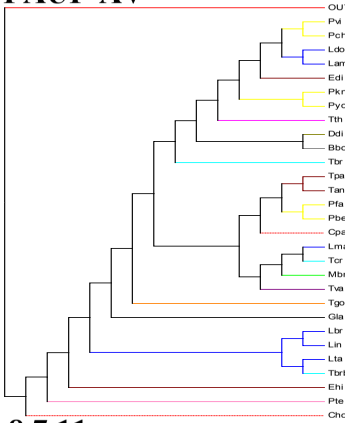
8.7.9

MRBAYES



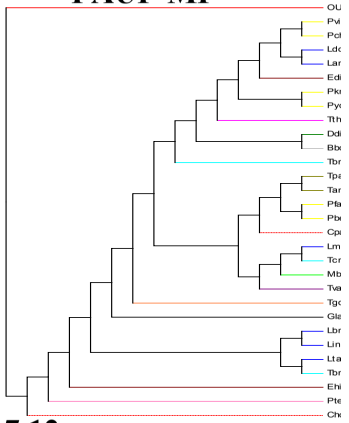
8.7.10

PAUP-AV



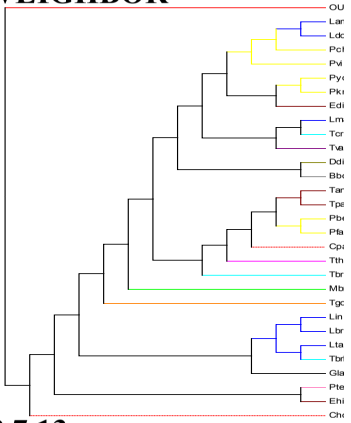
8.7.11

PAUP-MP



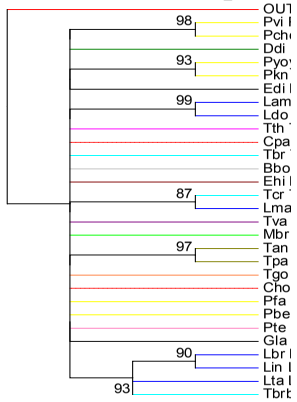
8.7.12

WEIGHBOR

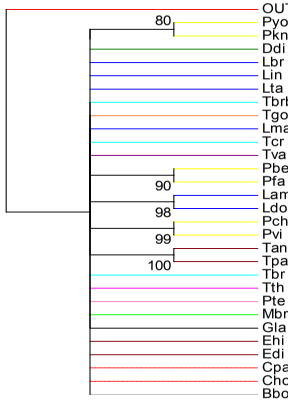


8.7.13

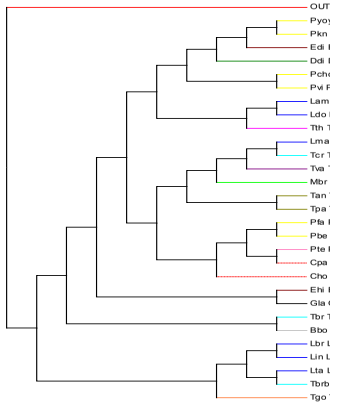
MRPA - Quatro Sequências Trimadas com o GBLOCKS



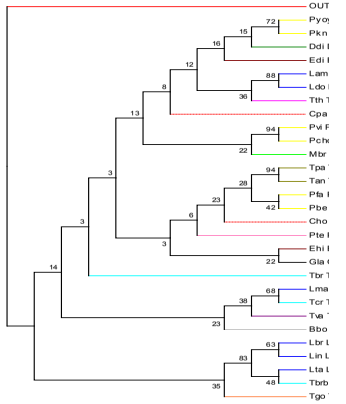
8.7.14



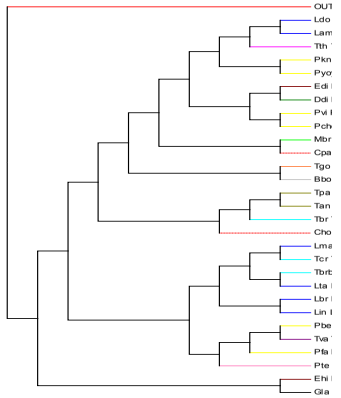
8.7.15



8.7.16

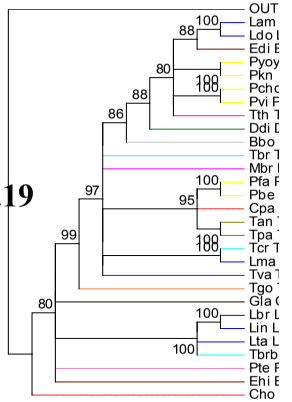


8.7.17

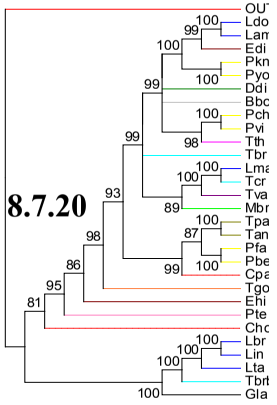


8.7.18

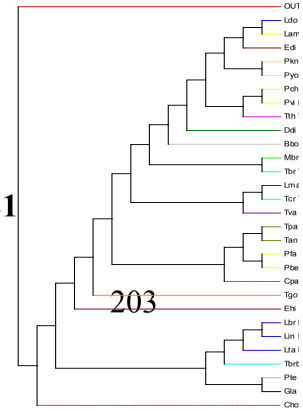
MRPA - Quatro Sequências Trimadas com o TRIMAL



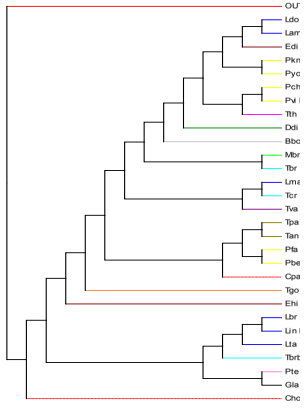
8.7.19



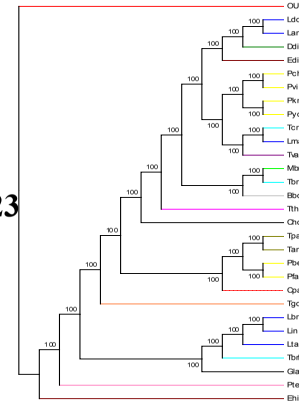
8.7.20



8.7.21

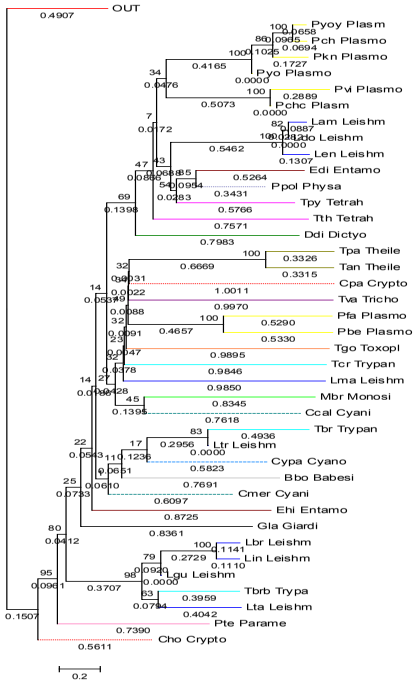


8.7.22



8.7.23

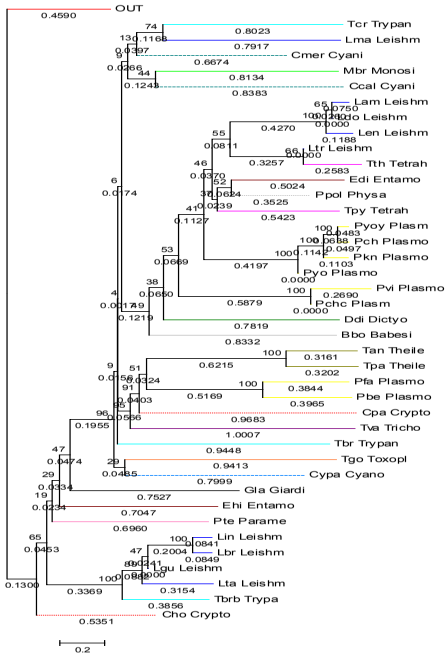
MRPA - Todas as Sequências Completas
Topologia com comprimento de ramo
PHYML



8.7.24

MRPA - Todas as Sequências Trimadas com o TRIMAL

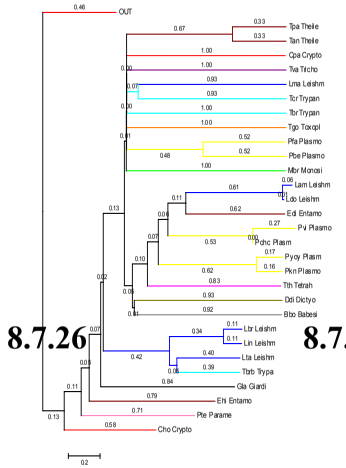
PHYML



8.7.25

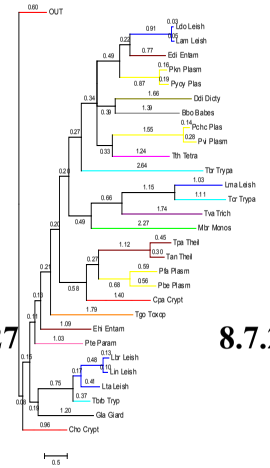
MRPA - Quatro Sequências Completas

PHYML



8.7.26

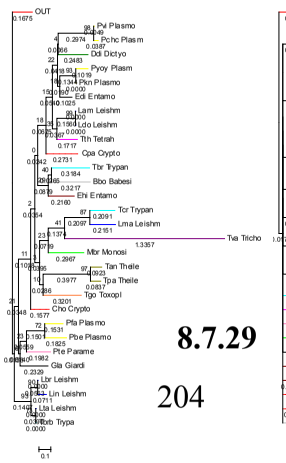
MRBAYES



8.7.27

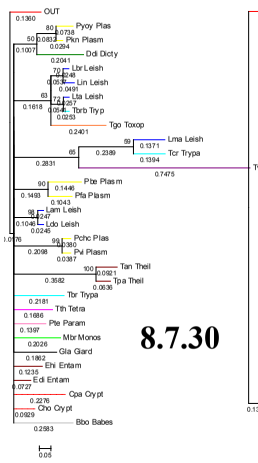
Trimadas com o GBLOCKS

PHYML



8.7.28

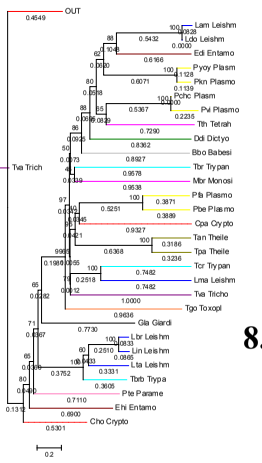
MRBAYES



8.7.29

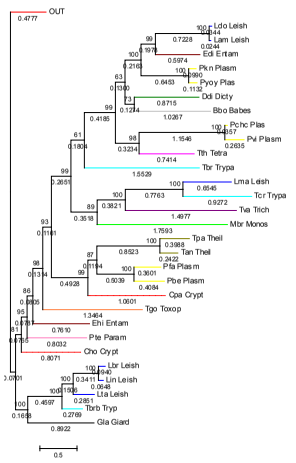
Trimadas com o TRIMAL

PHYML



8.7.30

MRBAYES



8.7.31

Anexo 8.7 - MRPA- Árvores filogenéticas construídas com os programas PHYML, MRBAYES, PAUP-AV, PAUP-MP e WEIGHBOR

As árvores filogenéticas do gene MRPA foram construídas segundo a Metodologia M2 para os genes candidatos de resistência às drogas.

Todas as Sequências Completas:
- a *árvore* 8.7.1 representa o PHYML,
- a *árvore* 8.7.2 representa o PAUP-AV,
- a *árvore* 8.7.3 representa o PAUP-MP,
- a *árvore* 8.7.4 representa o WEIGHBOR e
- a *árvore comprimento do ramo* 8.7.24 representa o PHYML.

Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 8.7.5 representa o PHYML,
- a *árvore* 8.7.6 representa o PAUP-AV,
- a *árvore* 8.7.7 representa o PAUP-MP,
- a *árvore* 8.7.8 representa o WEIGHBOR e
- a *árvore comprimento do ramo* 8.7.25 representa o PHYML.

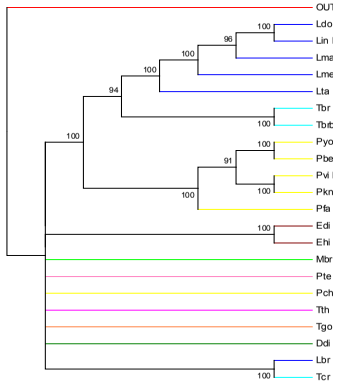
Quatro Sequências Completas:
- a *árvore* 8.7.9 representa o PHYML,
- a *árvore* 8.7.10 representa o MRBAYES,
- a *árvore* 8.7.11 representa o PAUP-AV,
- a *árvore* 8.7.12 representa o PAUP-MP,
- a *árvore* 8.7.13 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.7.26 representa o PHYML e
- a *árvore comprimento do ramo* 8.7.27 representa o MRBAYES.

Quatro Sequências Trimadas com o GBLOCKS:
- a *árvore* 8.7.14 representa o PHYML,
- a *árvore* 8.7.15 representa o MRBAYES,
- a *árvore* 8.7.16 representa o PAUP-AV,
- a *árvore* 8.7.17 representa o PAUP-MP,
- a *árvore* 8.7.18 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.7.28 representa o PHYML e
- a *árvore comprimento do ramo* 8.7.29 representa o MRBAYES.

Quatro Sequências Trimadas com o TRIMAL:
- a *árvore* 8.7.19 representa o PHYML,
- a *árvore* 8.7.20 representa o MRBAYES,
- a *árvore* 8.7.21 representa o PAUP-AV,
- a *árvore* 8.7.22 representa o PAUP-MP,
- a *árvore* 8.7.23 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.7.30 representa o PHYML e
- a *árvore comprimento do ramo* 8.7.31 representa o MRBAYES.

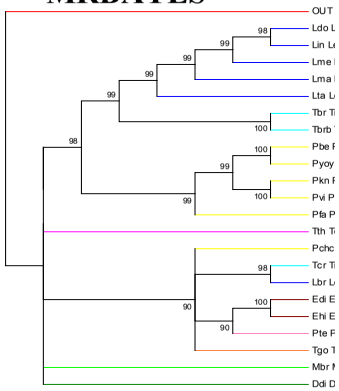
AQP - Todas as Sequências Completas
Mostrar só topologia

PHYML



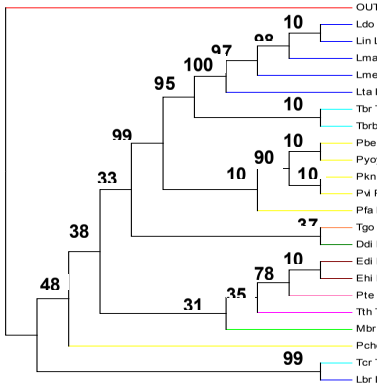
8.8.1

MRBAYES



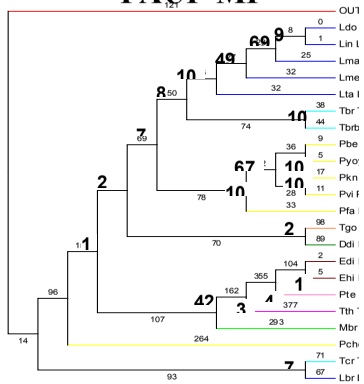
8.8.2

PAUP-AV



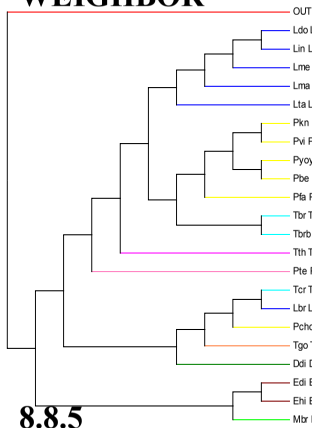
8.8.3

PAUP-MP



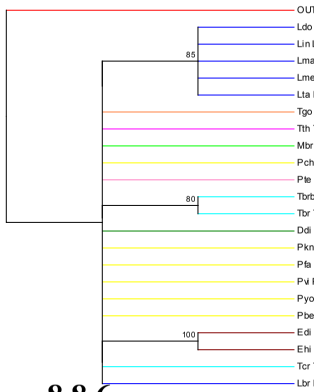
8.8.4

WEIGHBOR

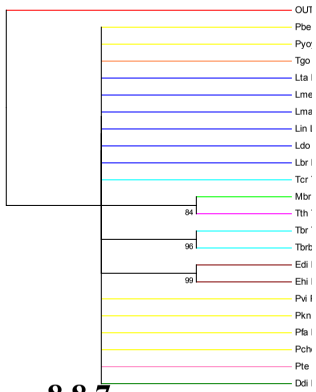


8.8.5

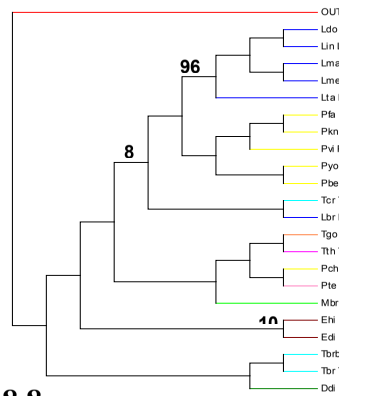
Trimadas com o GBLOCKS



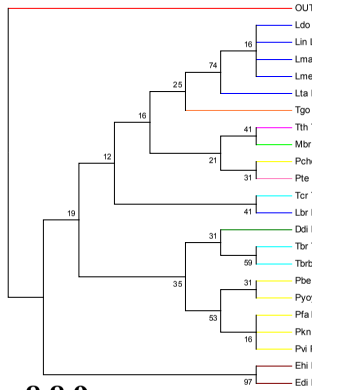
8.8.6



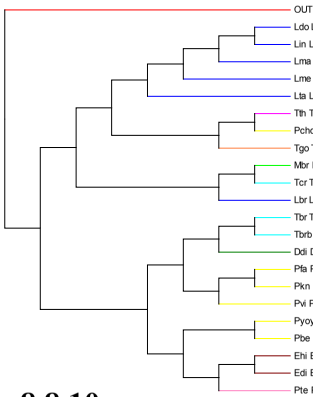
8.8.7



8.8.8

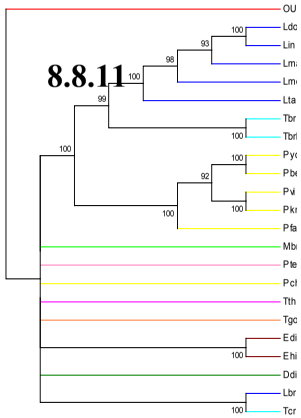


8.8.9

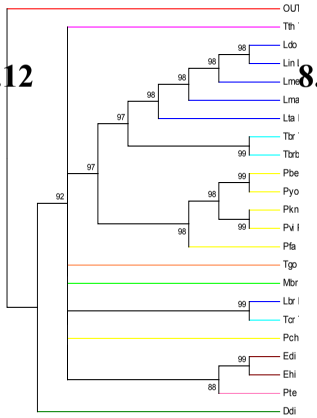


8.8.10

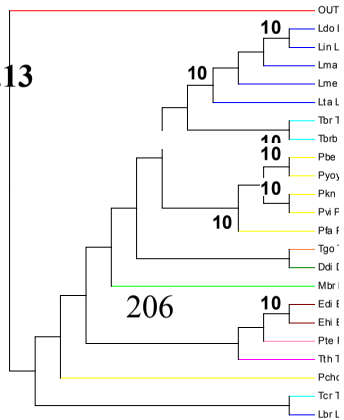
Trimadas com o TRIMAL



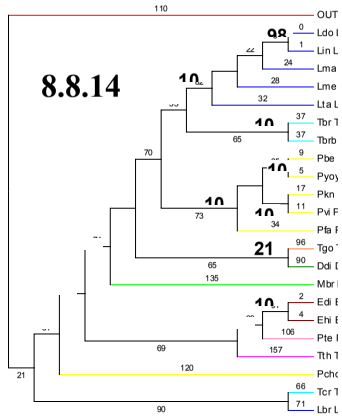
8.8.11



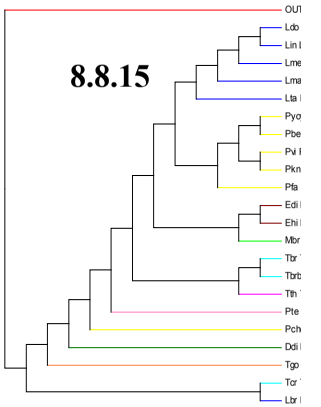
8.8.12



8.8.13



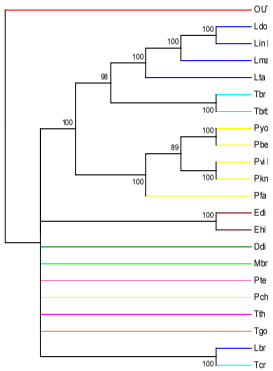
8.8.14



8.8.15

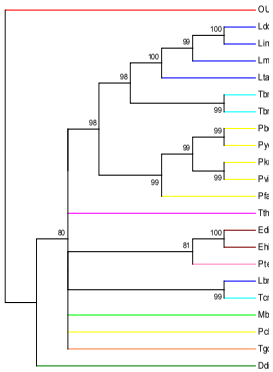
AQP - Quatro Sequências Completas
Mostrar só topologia

PHYML



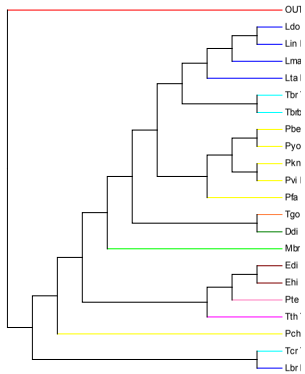
8.8.16

MRBAYES



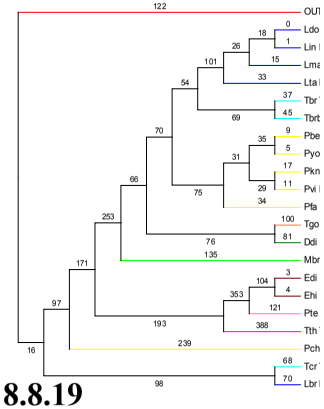
8.8.17

PAUP-AV



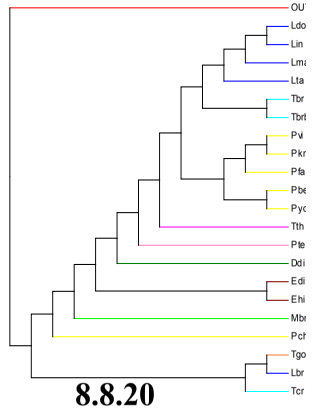
8.8.18

PAUP-MP



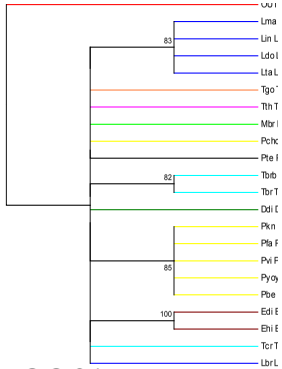
8.8.19

WEIGHBOR

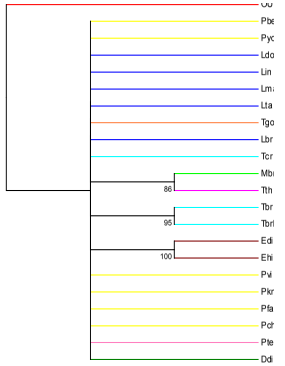


8.8.20

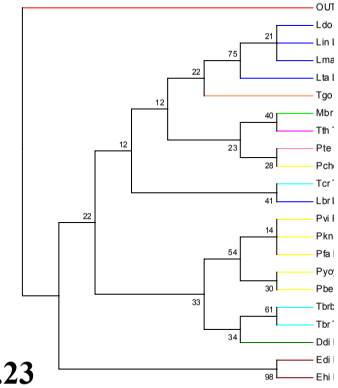
Trimadas com o GBLOCKS



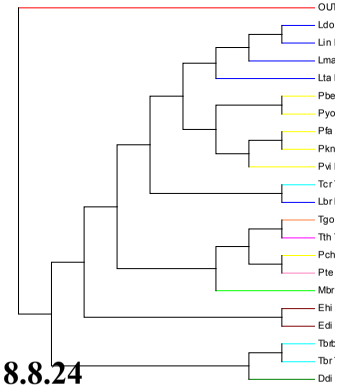
8.8.21



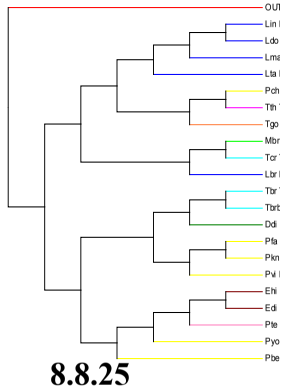
8.8.22



8.8.23

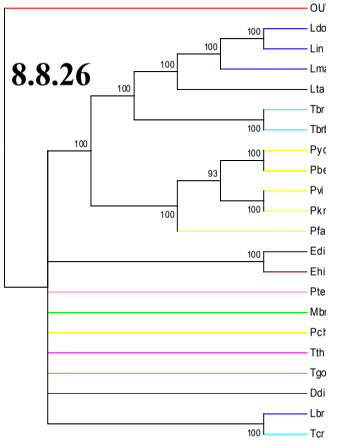


8.8.24

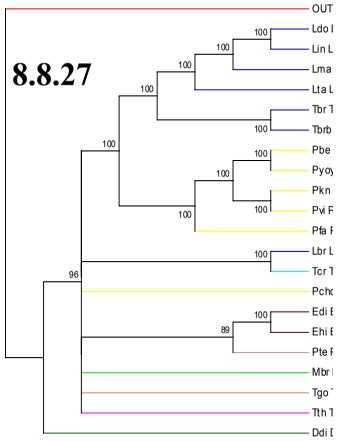


8.8.25

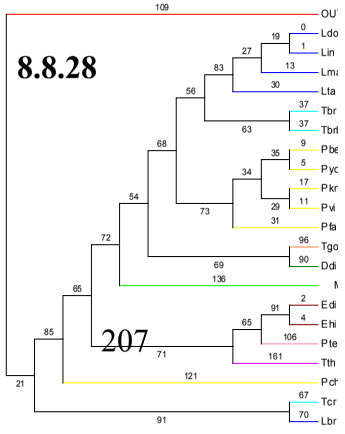
Trimadas com o TRIMAL



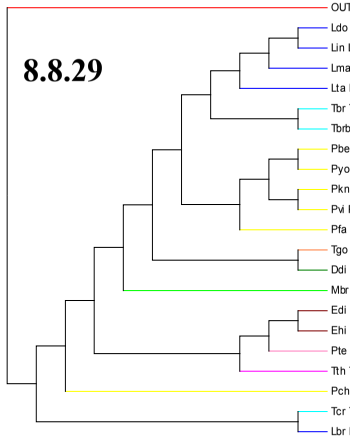
8.8.26



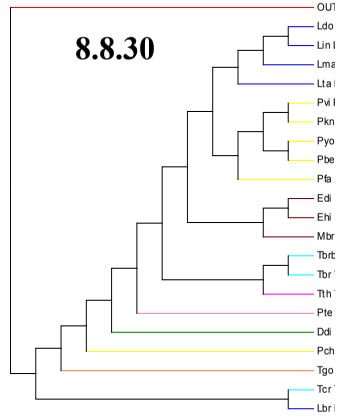
8.8.27



8.8.28



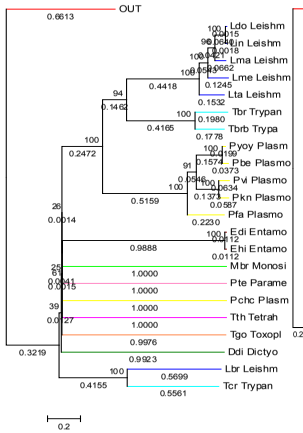
8.8.29



8.8.30

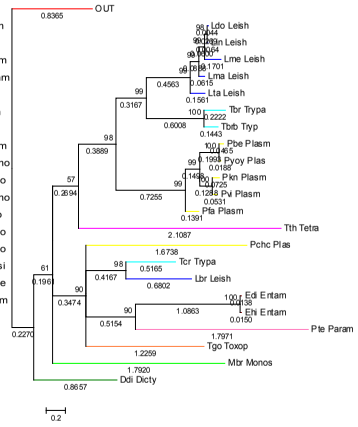
AQP - Todas as Sequências Completas Topologia com comprimento de ramo

PHYML



8.8.31

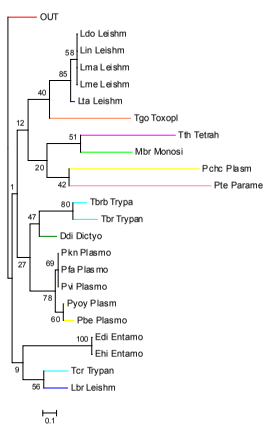
MRBAYES



8.8.32

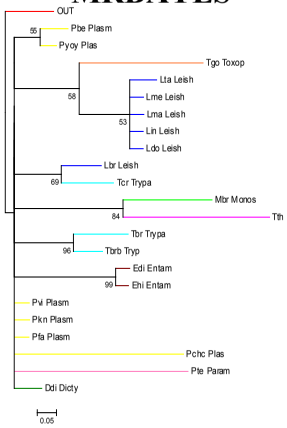
Trimadas com o GBLOCKS

PHYML



8.8.33

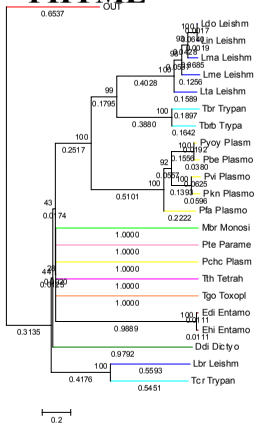
MRBAYES



8.8.34

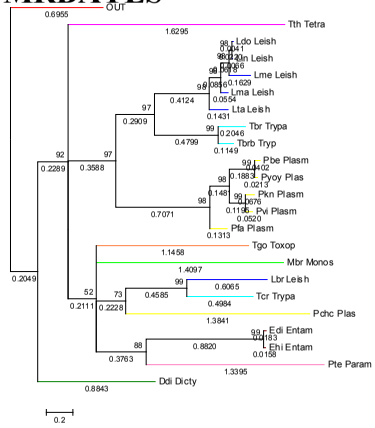
Trimadas com o TRIMAL

PHYML



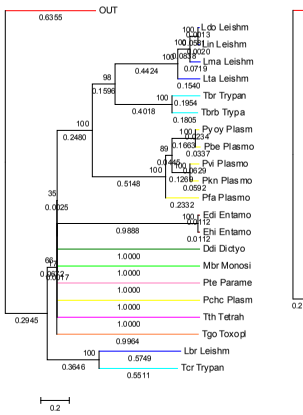
8.8.35

MRBAYES

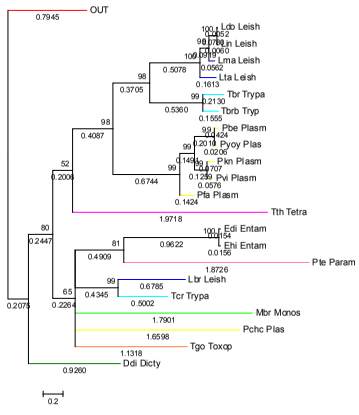


8.8.36

AQP - Quatro Sequências Completas

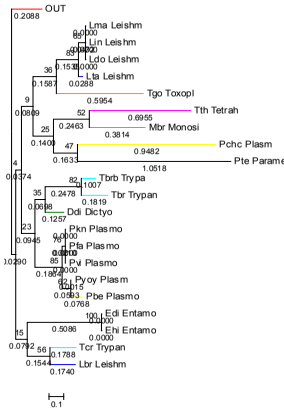


8.8.37

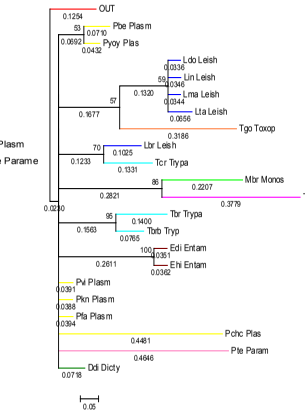


8.8.38

Trimadas com o GBLO CKS

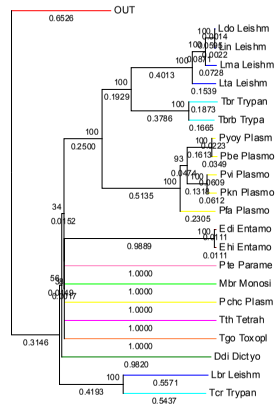


8.8.39

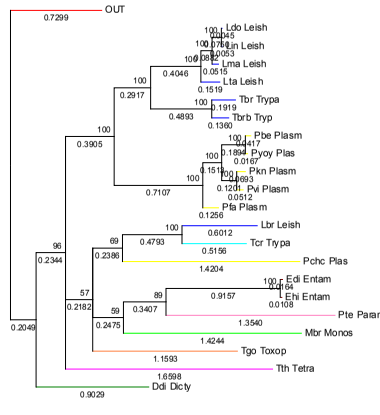


8.8.40

Trimadas com o TRIMAL



8.8.41



8.8.42

Anexo 8.8 - AQP - Árvores filogenéticas construídas com os programas PHYML, MRBAYES, PAUP-AV, PAUP-MP e WEIGHBOR

As árvores filogenéticas do gene AQP foram construídas segundo a Metodologia M2 para os genes candidatos de resistência às drogas.

Todas as Sequências Completas:
- a *árvore* 8.8.1 representa o PHYML,
- a *árvore* 8.8.2 representa o MRBAYES,
- a *árvore* 8.8.3 representa o PAUP-AV,
- a *árvore* 8.8.4 representa o PAUP-MP,
- a *árvore* 8.8.5 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.8.31 representa o PHYML e
- a *árvore comprimento do ramo* 8.8.32 representa o MRBAYES.

Todas as Sequências Trimadas com o GBLOCKS:
- a *árvore* 8.8.6 representa o PHYML,
- a *árvore* 8.8.7 representa o MRBAYES,
- a *árvore* 8.8.8 representa o PAUP-AV,
- a *árvore* 8.8.9 representa o PAUP-MP,
- a *árvore* 8.8.10 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.8.33 representa o PHYML e
- a *árvore comprimento do ramo* 8.8.34 representa o MRBAYES.

Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 8.8.11 representa o PHYML,
- a *árvore* 8.8.12 representa o MRBAYES,
- a *árvore* 8.8.13 representa o PAUP-AV,
- a *árvore* 8.8.14 representa o PAUP-MP,
- a *árvore* 8.8.15 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.8.35 representa o PHYML e
- a *árvore comprimento do ramo* 8.8.36 representa o MRBAYES.

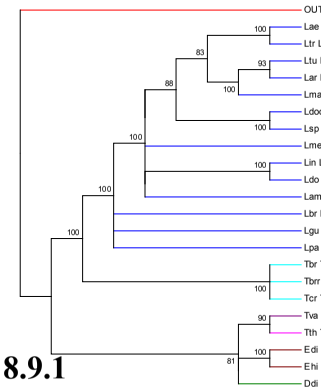
Quatro Sequências Completas:
- a *árvore* 8.8.16 representa o PHYML,
- a *árvore* 8.8.17 representa o MRBAYES,
- a *árvore* 8.8.18 representa o PAUP-AV,
- a *árvore* 8.8.19 representa o PAUP-MP,
- a *árvore* 8.8.20 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.8.37 representa o PHYML e
- a *árvore comprimento do ramo* 8.8.38 representa o MRBAYES.

Quatro Sequências Trimadas com o GBLOCKS:
- a *árvore* 8.8.21 representa o PHYML,
- a *árvore* 8.8.22 representa o MRBAYES,
- a *árvore* 8.8.23 representa o PAUP-AV,
- a *árvore* 8.8.24 representa o PAUP-MP,
- a *árvore* 8.8.25 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.8.39 representa o PHYML e
- a *árvore comprimento do ramo* 8.8.40 representa o MRBAYES.

Quatro Sequências Trimadas com o TRIMAL:
- a *árvore* 8.8.26 representa o PHYML,
- a *árvore* 8.8.27 representa o MRBAYES,
- a *árvore* 8.8.28 representa o PAUP-AV,
- a *árvore* 8.8.29 representa o PAUP-MP,
- a *árvore* 8.8.30 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.8.41 representa o PHYML e
- a *árvore comprimento do ramo* 8.8.42 representa o MRBAYES.

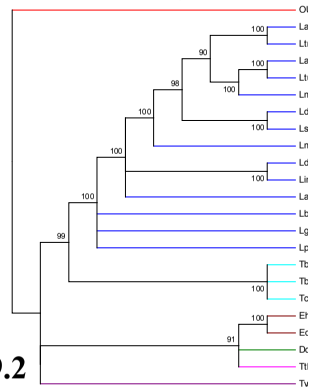
GP63 - Todas as Sequências Completas
Mostrar só topologia

PHYML



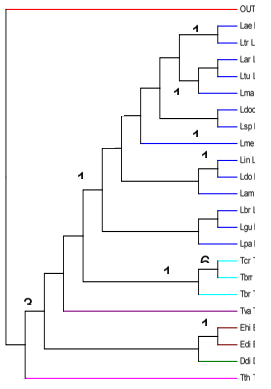
8.9.1

MRBAYES



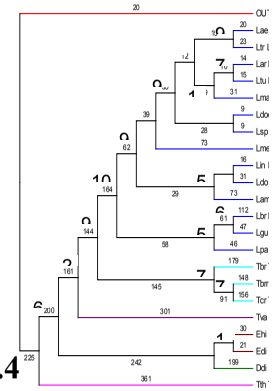
8.9.2

PAUP-AV



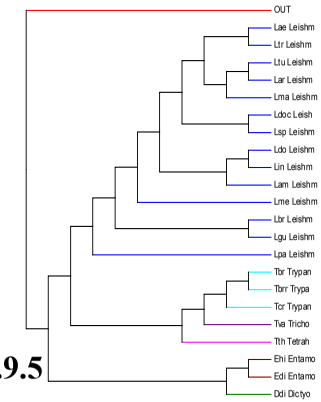
8.9.3

PAUP-MP



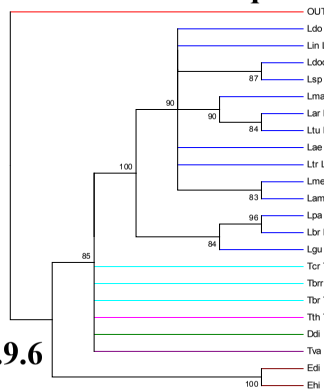
8.9.4

WEIGHBOR

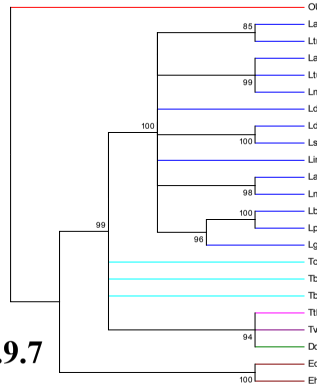


8.9.5

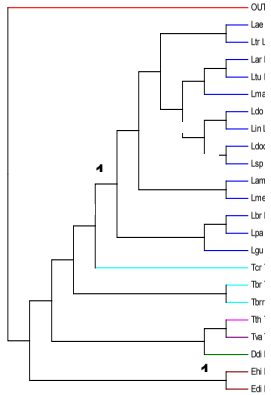
GP63 - Todas as Sequências Trimadas com o GBLOCKS



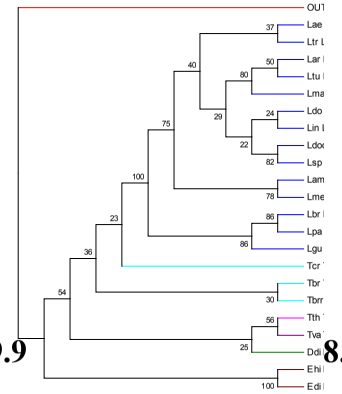
8.9.6



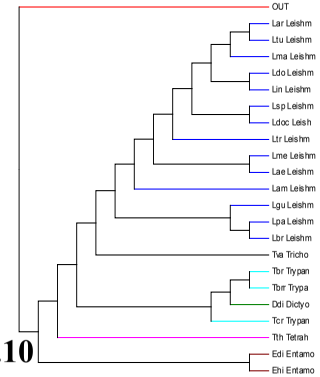
8.9.7



8.9.8

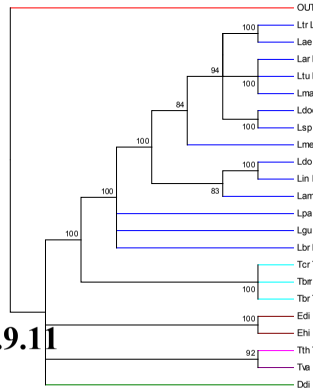


8.9.9

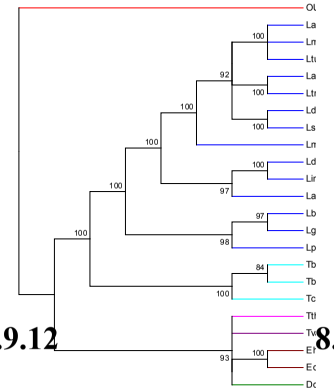


8.9.10

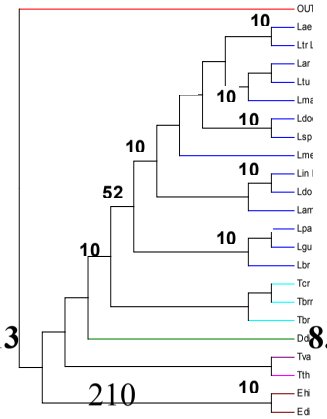
GP63 - Todas as Sequências Trimadas com o TRIMAL



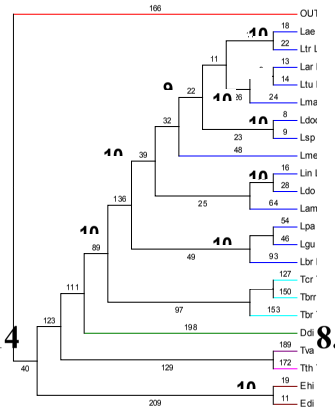
8.9.11



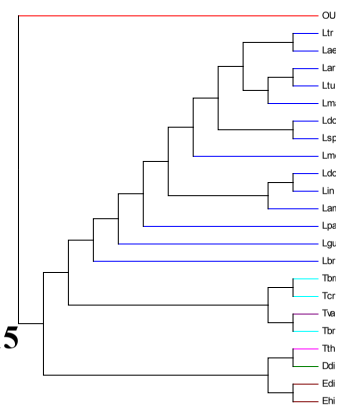
8.9.12



8.9.13



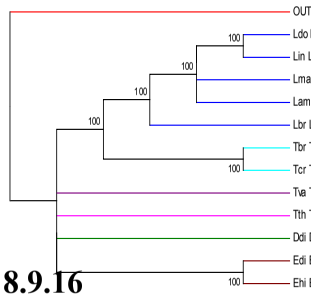
8.9.14



8.9.15

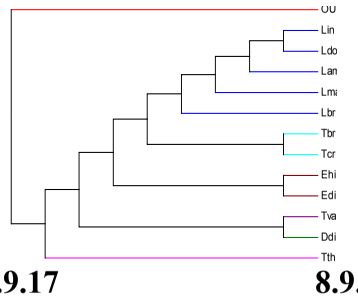
GP63 - Quatro Sequências Completas
Mostrar só topologia

PHYML



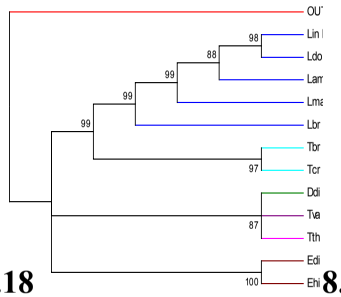
8.9.16

MRBAYES



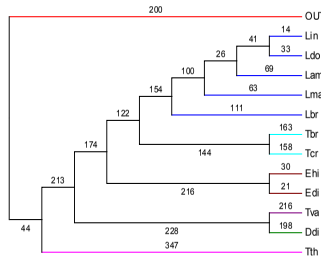
8.9.17

PAUP-AV



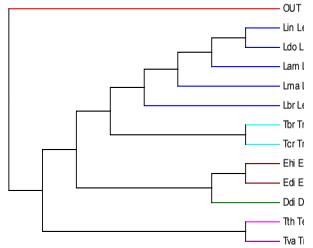
8.9.18

PAUP-MP



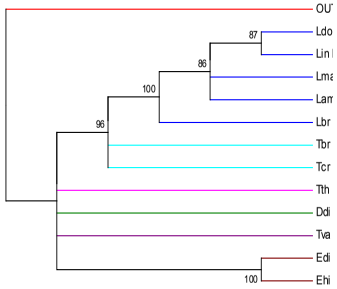
8.9.19

WEIGHBOR

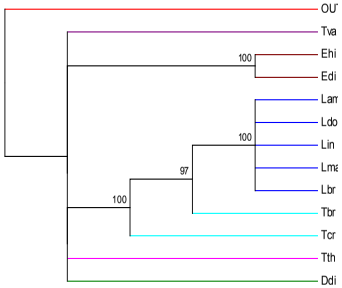


8.9.20

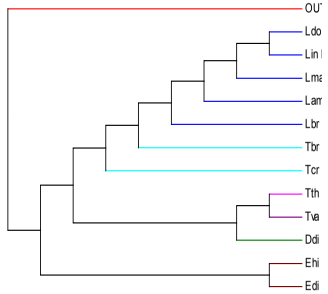
GP63 - Quatro Sequências Trimadas com o GBLOCKS



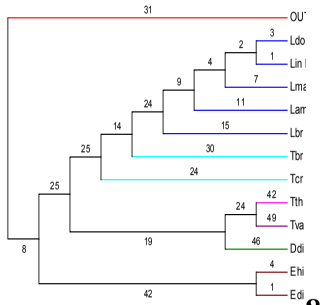
8.9.21



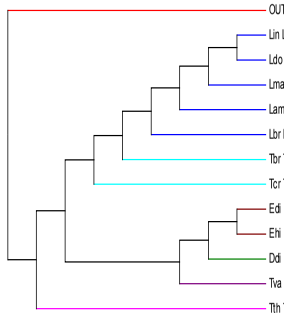
8.9.22



8.9.23

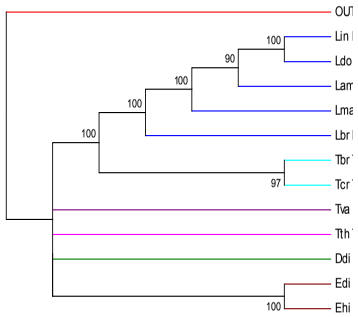


8.9.24

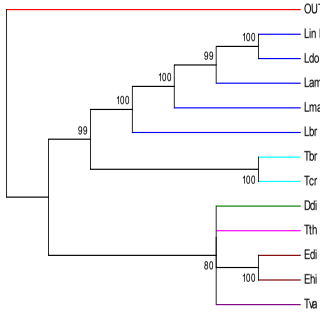


8.9.25

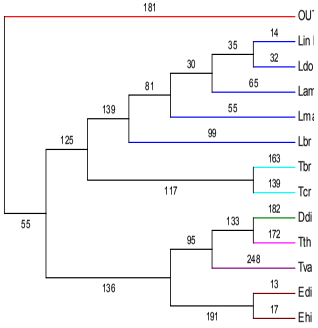
GP63 - Quatro Sequências Trimadas com o TRIMAL



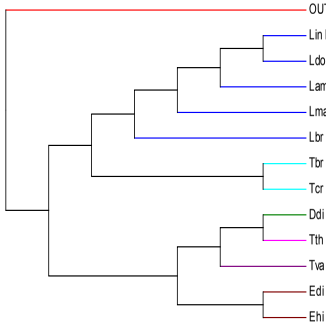
8.9.26



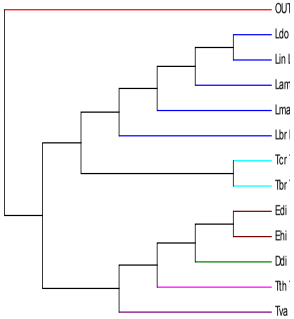
8.9.27



8.9.28



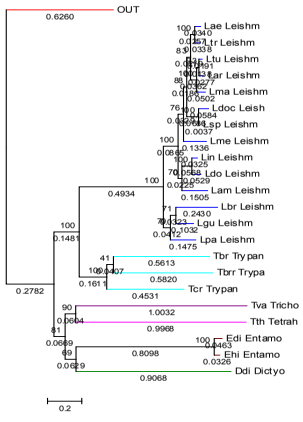
8.9.29



8.9.30

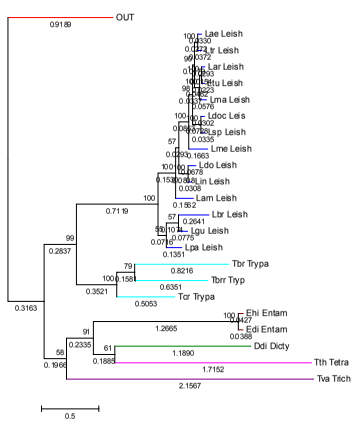
GP63 - Todas as Sequências Completas
Topologia com comprimento de ramo

PHYML



8.9.31

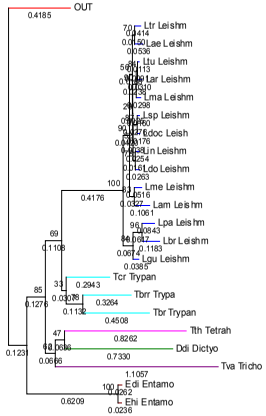
MRBAYES



8.9.32

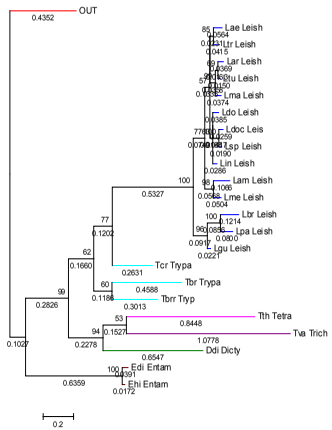
Trimadas com o GBLOCKS

PHYML



8.9.33

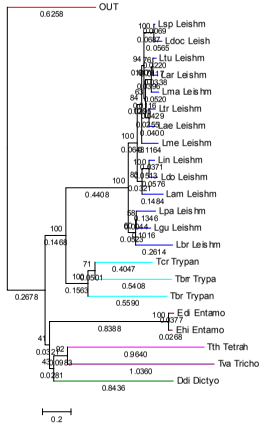
MRBAYES



8.9.34

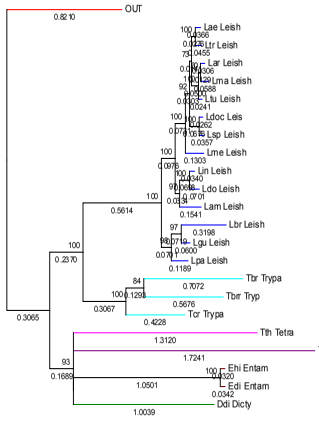
Trimadas com o TRIMAL

PHYML



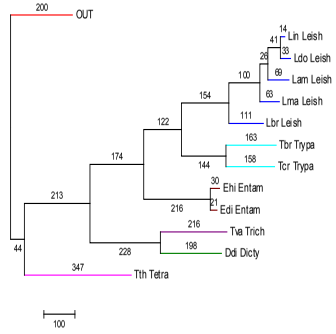
8.9.35

MRBAYES

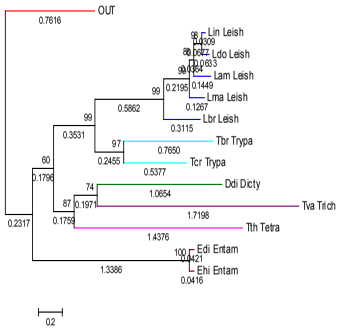


8.9.36

GP63 - Quatro Sequências Completas

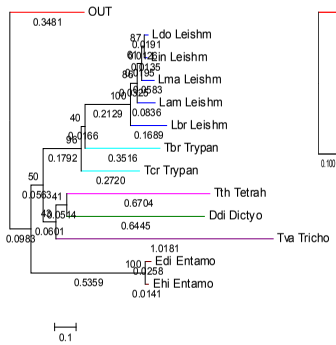


8.9.37

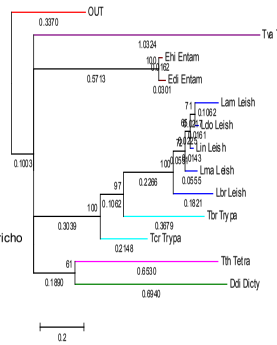


8.9.38

Trimadas com o GBLOCKS

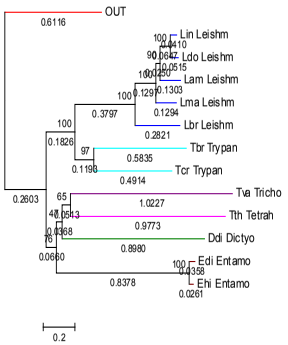


8.9.39

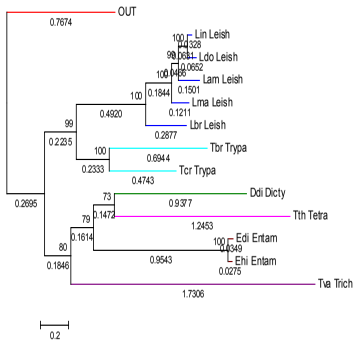


8.9.40

Trimadas com o TRIMAL



8.9.41



8.9.42

Anexo 8.9 - GP63 - Árvores filogenéticas construídas com os programas PHYML, MRBAYES, PAUP-AV, PAUP-MP e WEIGHBOR

As árvores filogenéticas do gene GP63 foram construídas segundo a Metodologia M2 para os genes candidatos de resistência às drogas.

Todas as Sequências Completas:
- a *árvore* 8.9.1 representa o PHYML,
- a *árvore* 8.9.2 representa o MRBAYES,
- a *árvore* 8.9.3 representa o PAUP-AV,
- a *árvore* 8.9.4 representa o PAUP-MP,
- a *árvore* 8.9.5 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.9.31 representa o PHYML e
- a *árvore comprimento do ramo* 8.9.32 representa o MRBAYES.

Todas as Sequências Trimadas com o GBLOCKS:
- a *árvore* 8.9.6 representa o PHYML,
- a *árvore* 8.9.7 representa o MRBAYES,
- a *árvore* 8.9.8 representa o PAUP-AV,
- a *árvore* 8.9.9 representa o PAUP-MP,
- a *árvore* 8.9.10 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.9.33 representa o PHYML e
- a *árvore comprimento do ramo* 8.9.34 representa o MRBAYES.

Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 8.9.11 representa o PHYML,
- a *árvore* 8.9.12 representa o MRBAYES,
- a *árvore* 8.9.13 representa o PAUP-AV,
- a *árvore* 8.9.14 representa o PAUP-MP,
- a *árvore* 8.9.15 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.9.35 representa o PHYML e
- a *árvore comprimento do ramo* 8.9.36 representa o MRBAYES.

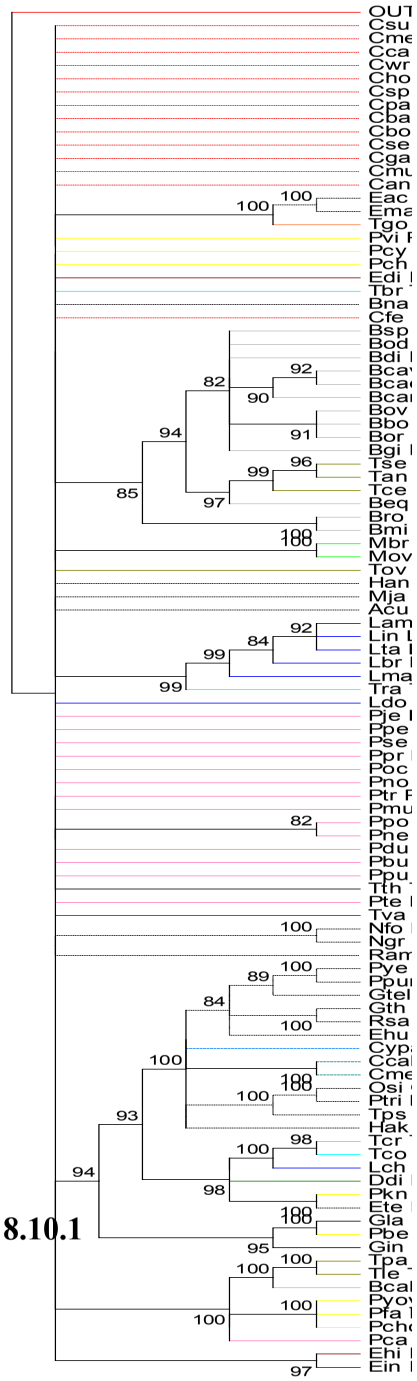
Quatro Sequências Completas:
- a *árvore* 8.9.16 representa o PHYML,
- a *árvore* 8.9.17 representa o MRBAYES,
- a *árvore* 8.9.18 representa o PAUP-AV,
- a *árvore* 8.9.19 representa o PAUP-MP,
- a *árvore* 8.9.20 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.9.37 representa o PHYML e
- a *árvore comprimento do ramo* 8.9.38 representa o MRBAYES.

Quatro Sequências Trimadas com o GBLOCKS:
- a *árvore* 8.9.21 representa o PHYML,
- a *árvore* 8.9.22 representa o MRBAYES,
- a *árvore* 8.9.23 representa o PAUP-AV,
- a *árvore* 8.9.24 representa o PAUP-MP,
- a *árvore* 8.9.25 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.9.39 representa o PHYML e
- a *árvore comprimento do ramo* 8.9.40 representa o MRBAYES.

Quatro Sequências Trimadas com o TRIMAL:
- a *árvore* 8.9.26 representa o PHYML,
- a *árvore* 8.9.27 representa o MRBAYES,
- a *árvore* 8.9.28 representa o PAUP-AV,
- a *árvore* 8.9.29 representa o PAUP-MP,
- a *árvore* 8.9.30 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.9.41 representa o PHYML e
- a *árvore comprimento do ramo* 8.9.42 representa o MRBAYES.

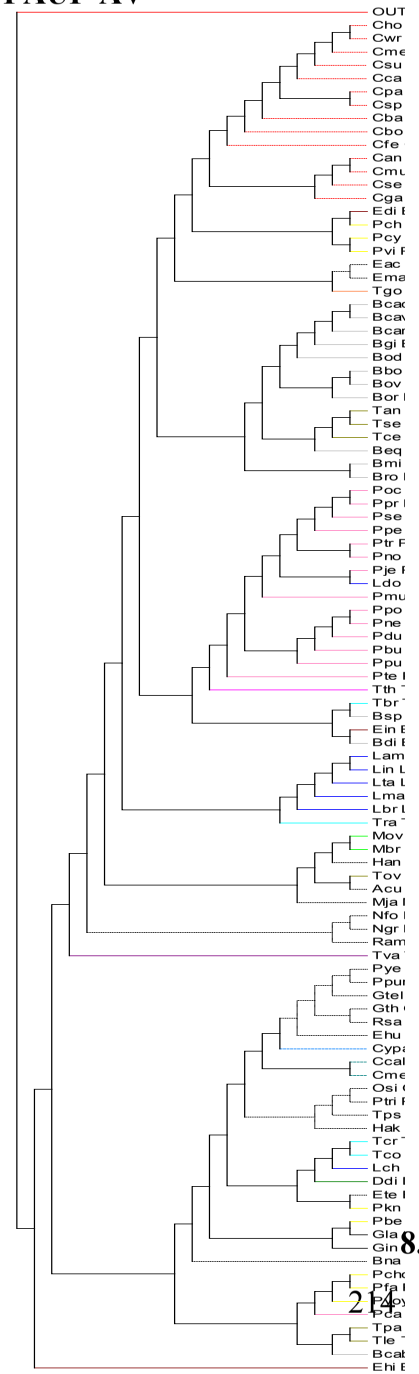
hsp70 - Todas as Sequências Completas

PHYML



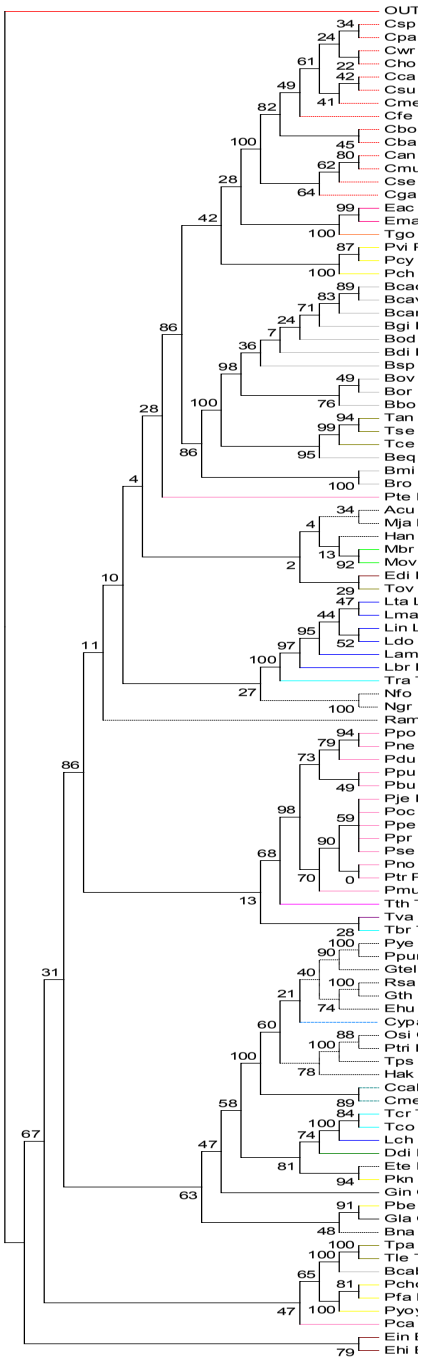
8.10.1

PAUP-AV



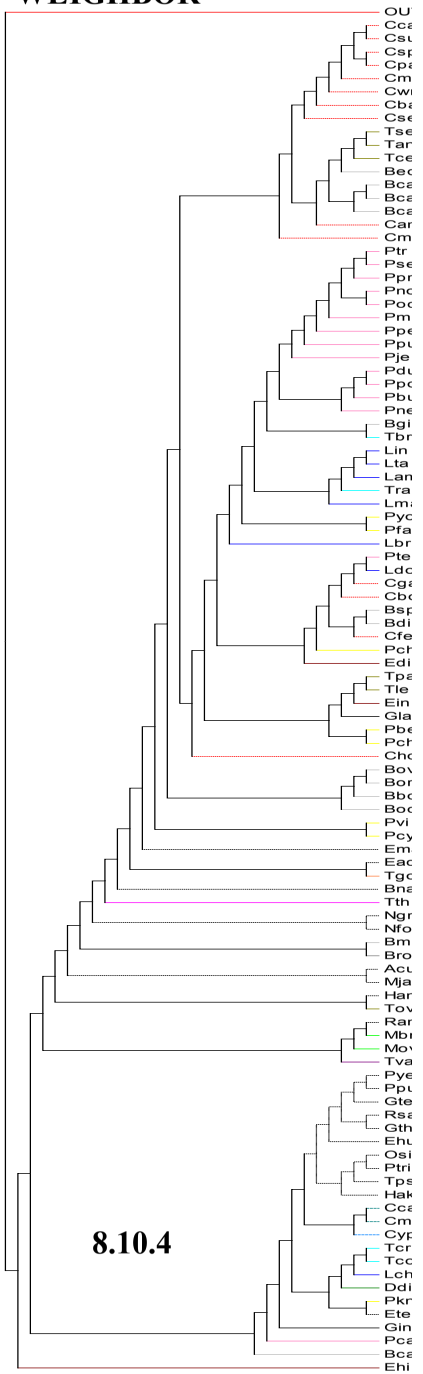
8.10.2

PAUP-MP



8.10.3

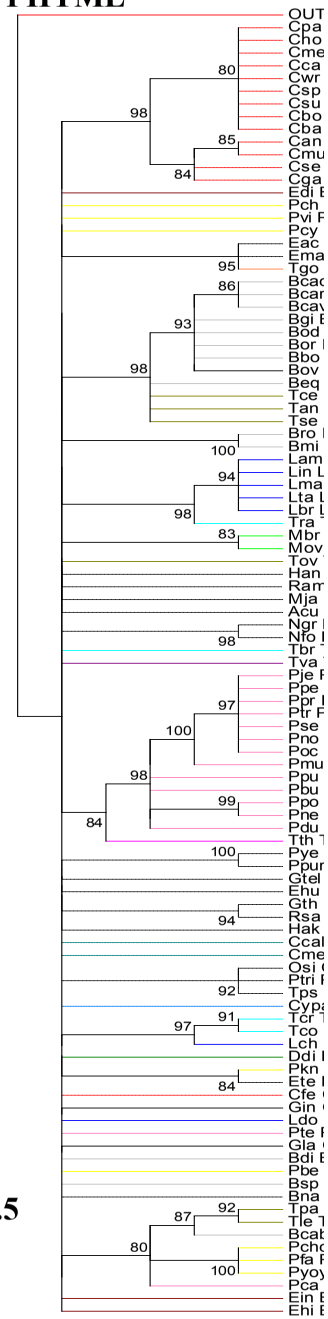
WEIGHBOR



8.10.4

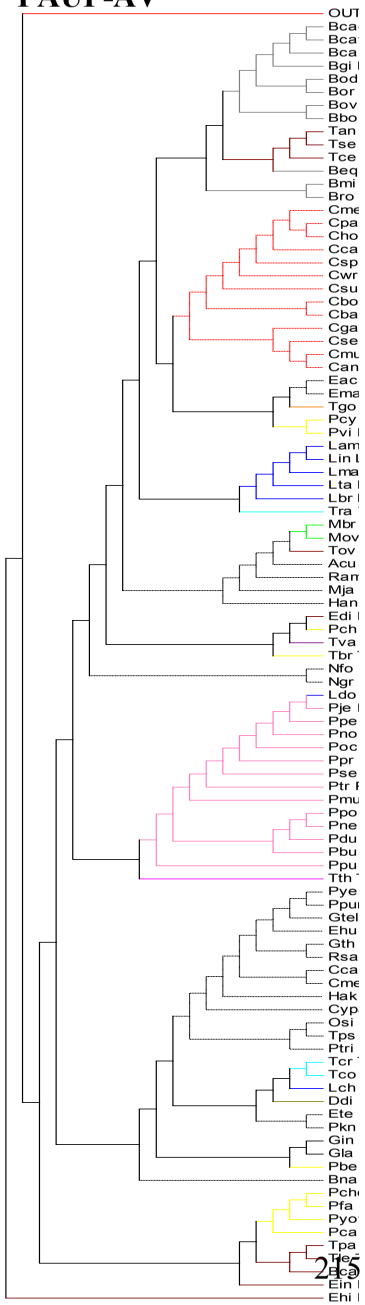
hsp70 - Todas as Sequências Trimadas com o TRIMAL
Mostrar só topologia

PHYML



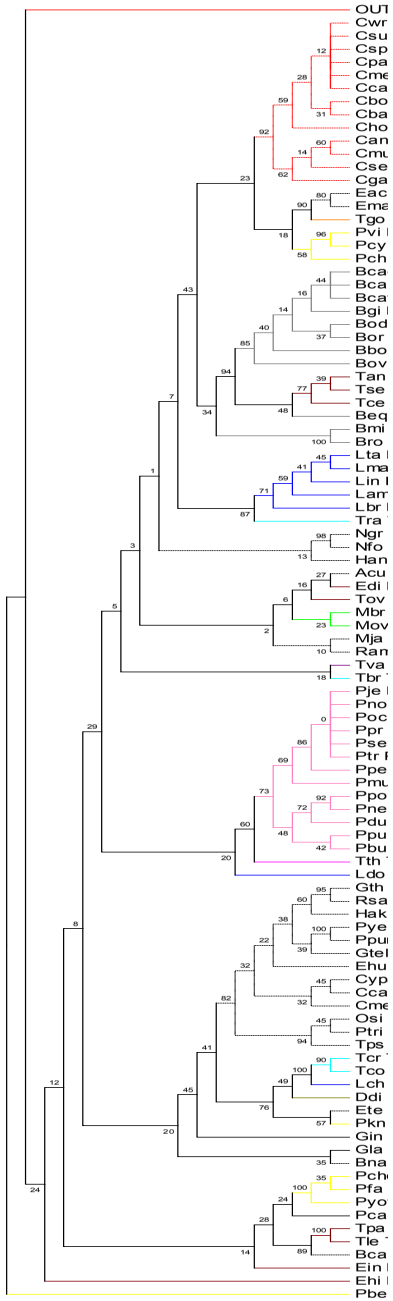
8.10.5

PAUP-AV



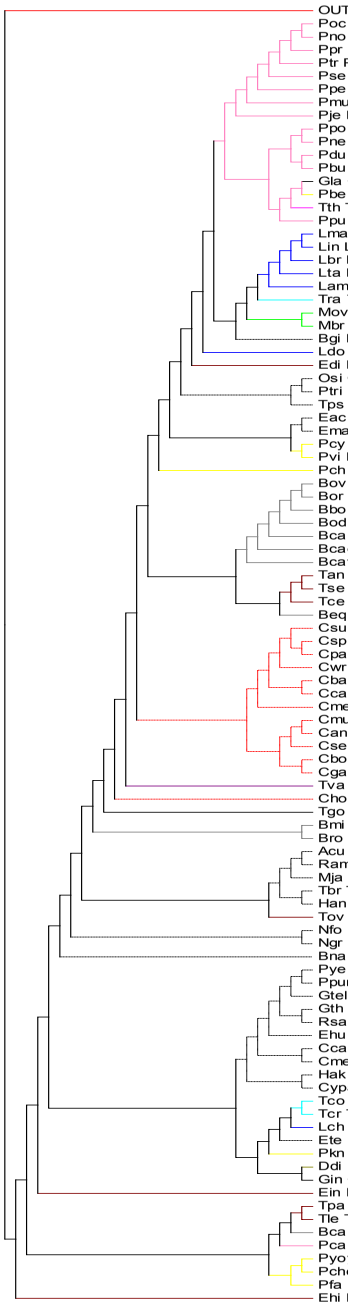
8.10.6

PAUP-MP



8.10.7

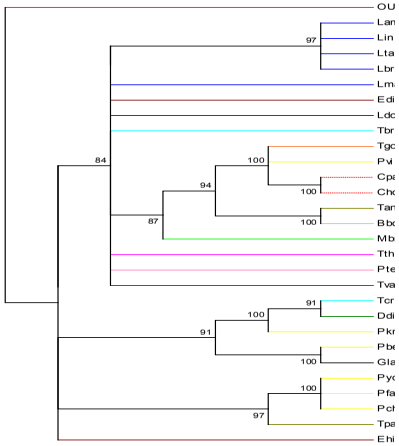
WEIGHBOR



8.10.8

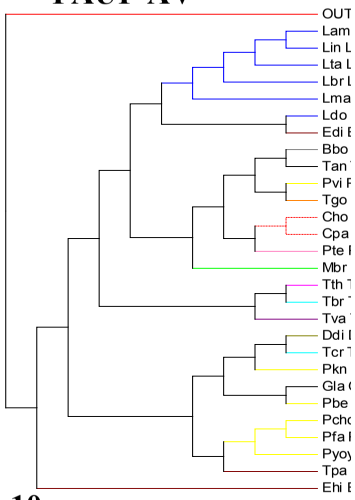
hsp70 - Quatro Sequências Completas
Mostrar só topologia

PHYML



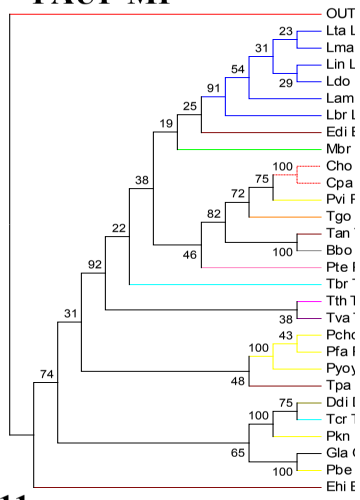
8.10.9

PAUP-AV



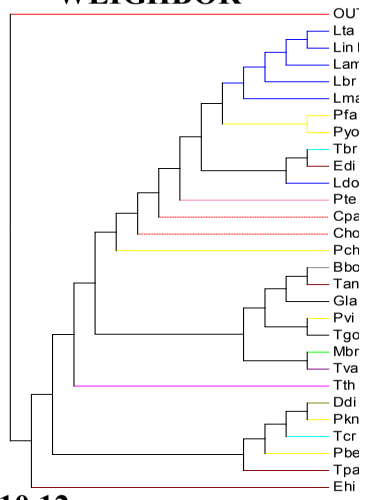
8.10.10

PAUP-MP



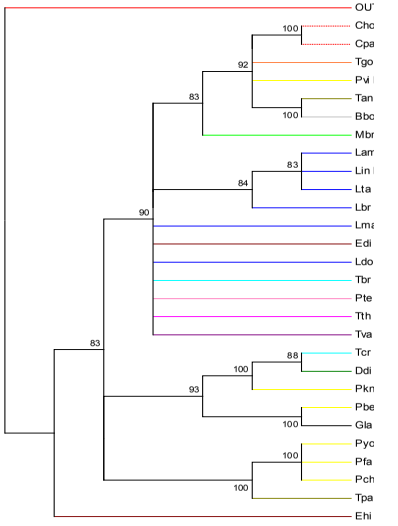
8.10.11

WEIGHBOR

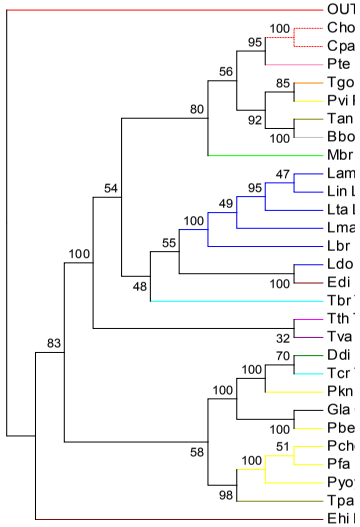


8.10.12

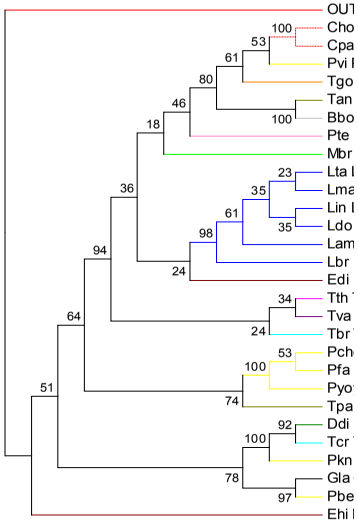
hsp70 - Quatro Sequências Trimadas com o TRIMAL



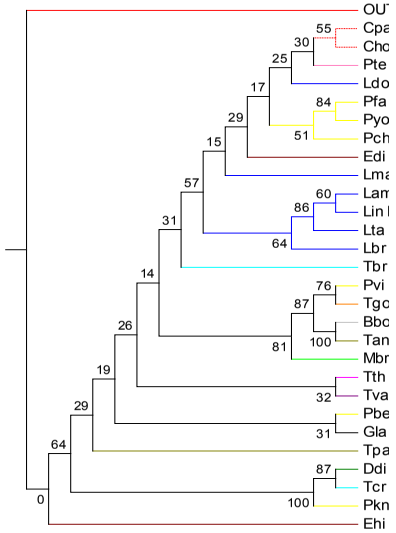
8.10.13



8.10.14



8.10.15

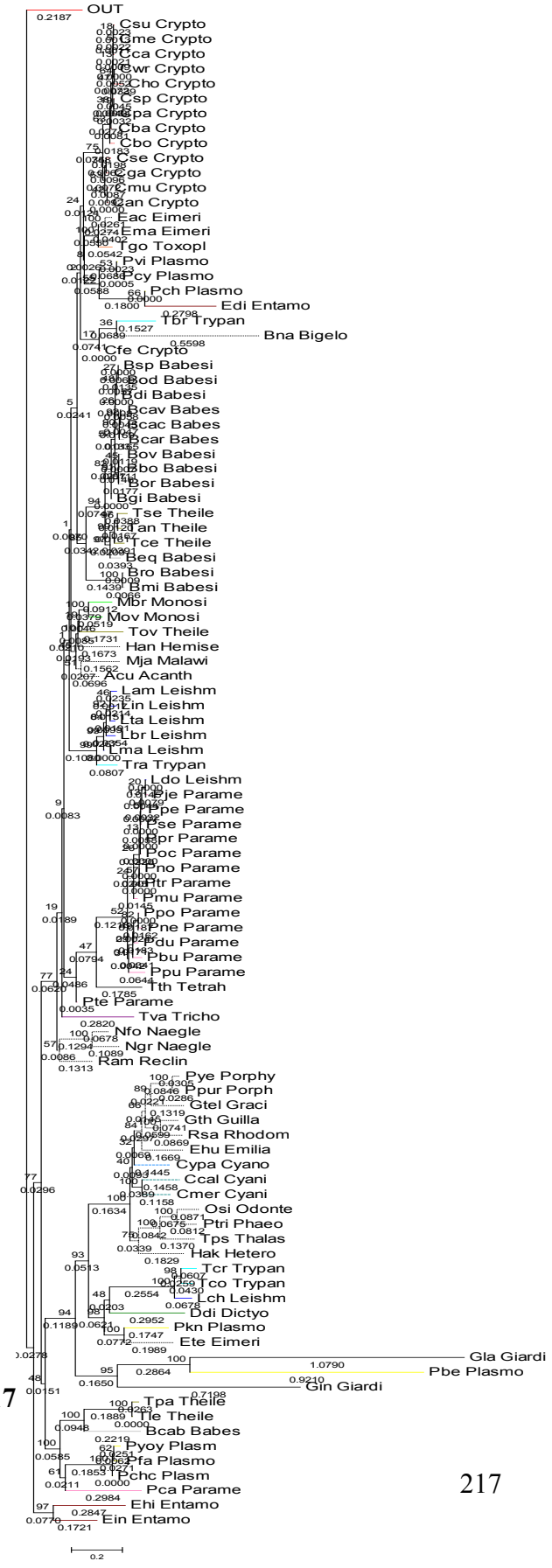


8.10.16

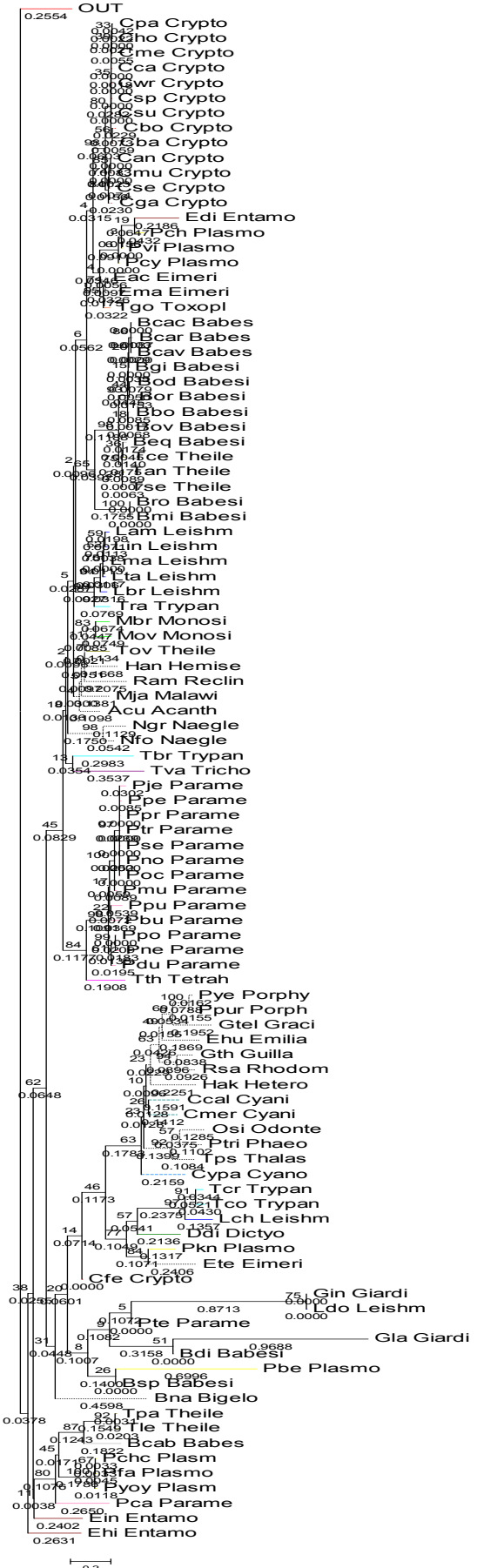
hsp70 - Todas as Sequências Completas

Trimadas com o TRIMAL

Topologia com comprimento de ramo - PHYML

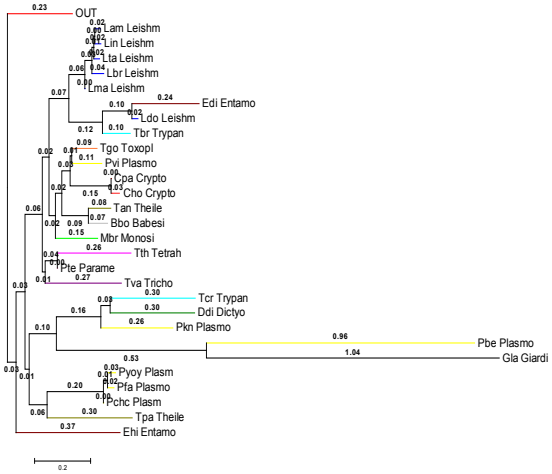


8.10.17



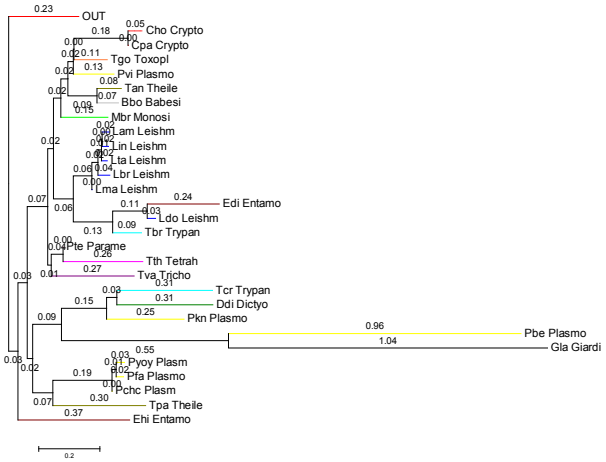
8.10.18

hsp70 - Quatro Sequências Completas Topologia com comprimento de ramo PHYML



8.10.19

Trimadas com o TRIMAL



8.10.20

Anexo 8.10 - hsp70 - Árvores filogenéticas construídas com os programas PHYML, MRBAYES, PAUP-AV, PAUP-MP e WEIGHBOR

As árvores filogenéticas do gene hsp70 foram construídas segundo a Metodologia M2 para os genes candidatos de resistência às drogas.

Todas as Sequências Completas:
- a *árvore* 8.10.1 representa o PHYML,
- a *árvore* 8.10.2 representa o PAUP-AV,
- a *árvore* 8.10.3 representa o PAUP-MP,
- a *árvore* 8.10.4 representa o WEIGHBOR e
- a *árvore comprimento do ramo* 8.10.17 representa o PHYML.

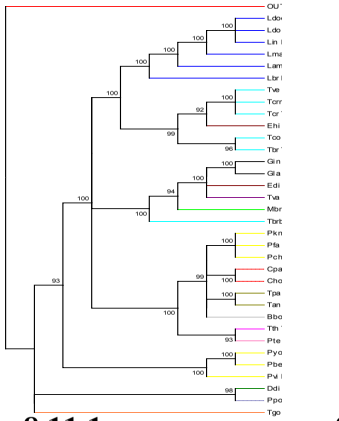
Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 8.10.5 representa o PHYML,
- a *árvore* 8.10.6 representa o PAUP-AV,
- a *árvore* 8.10.7 representa o PAUP-MP,
- a *árvore* 8.10.8 representa o WEIGHBOR e
- a *árvore comprimento do ramo* 8.10.18 representa o PHYML.

Quatro Sequências Completas:
- a *árvore* 8.10.9 representa o PHYML,
- a *árvore* 8.10.10 representa o PAUP-AV,
- a *árvore* 8.10.11 representa o PAUP-MP,
- a *árvore* 8.10.12 representa o WEIGHBOR e
- a *árvore comprimento do ramo* 8.10.19 representa o PHYML.

Quatro Sequências Trimadas com o TRIMAL:
- a *árvore* 8.10.13 representa o PHYML,
- a *árvore* 8.10.14 representa o PAUP-AV,
- a *árvore* 8.10.15 representa o PAUP-MP,
- a *árvore* 8.10.16 representa o WEIGHBOR e
- a *árvore comprimento do ramo* 8.10.20 representa o PHYML.

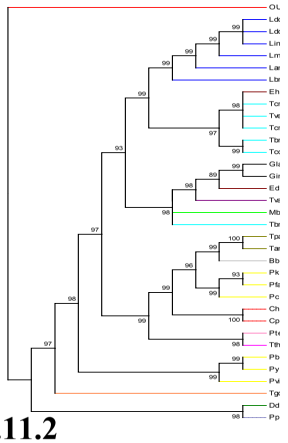
TRYR - Todas as Sequências Completas - Mostrar só topologia

PHYML



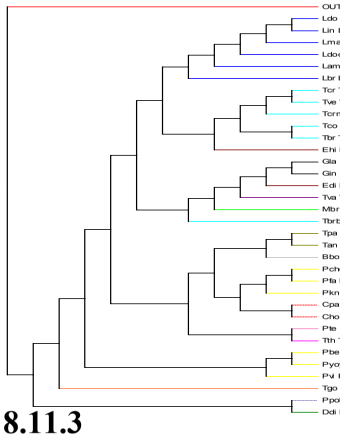
8.11.1

MRBAYES



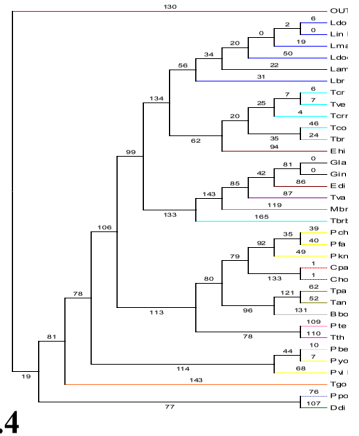
8.11.2

PAUP-AV



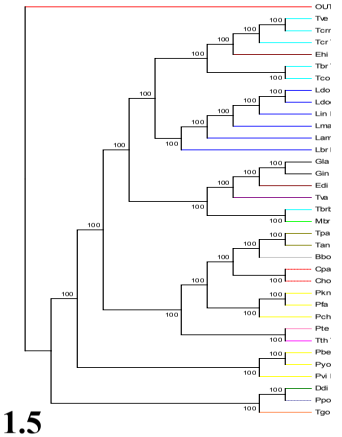
8.11.3

PAUP-MP



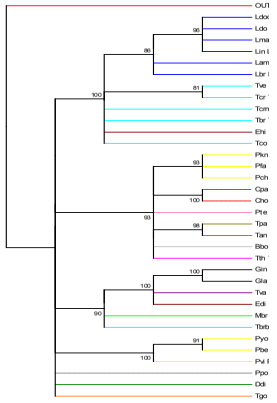
8.11.4

WEIGHBOR

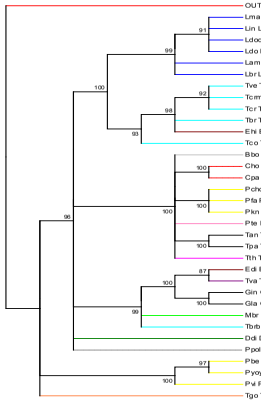


8.11.5

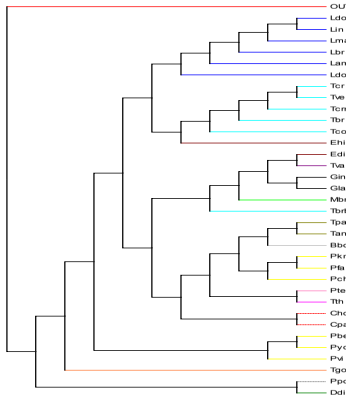
TRYR - Todas as Sequências Trimadas com o GBLOCKS



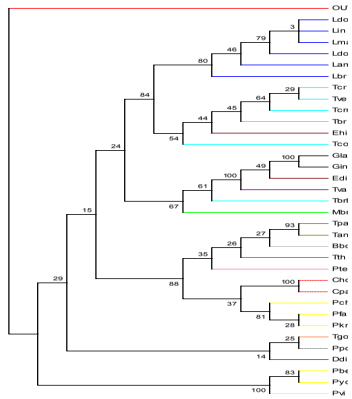
8.11.6



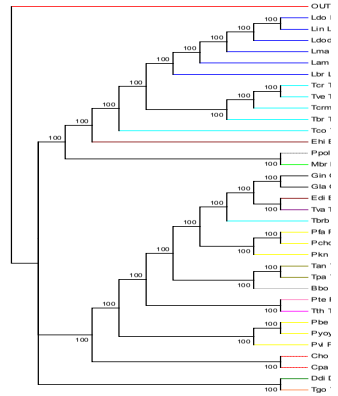
8.11.7



8.11.8

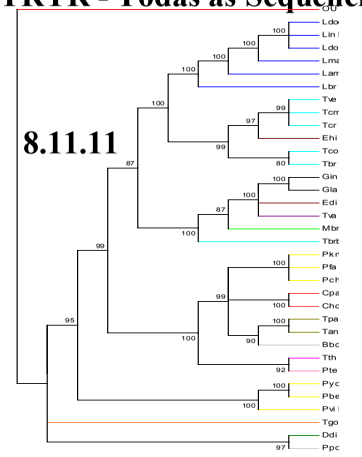


8.11.9

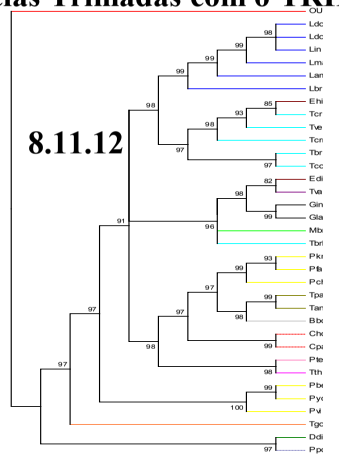


8.11.10

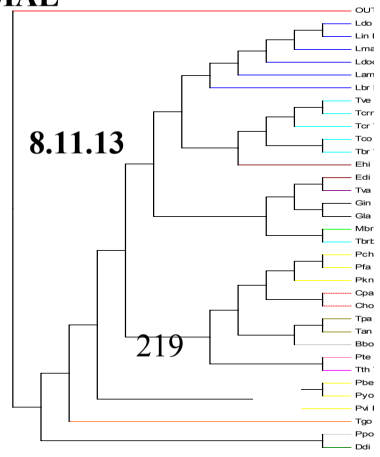
TRYR - Todas as Sequências Trimadas com o TRIMAL



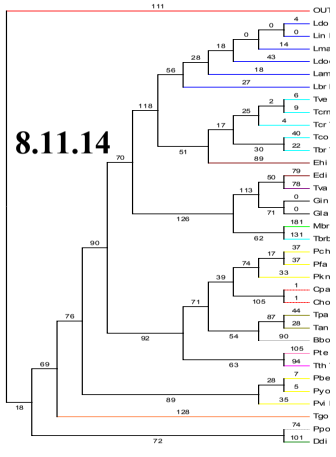
8.11.11



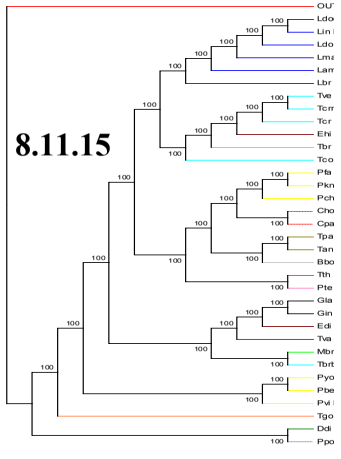
8.11.12



8.11.13



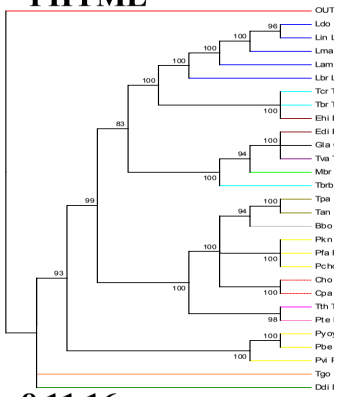
8.11.14



8.11.15

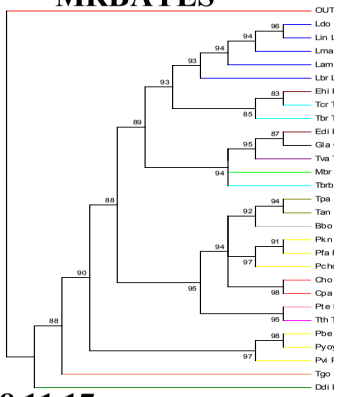
TRYR - Quatro as Sequências Completas - Mostrar só topologia

PHYML



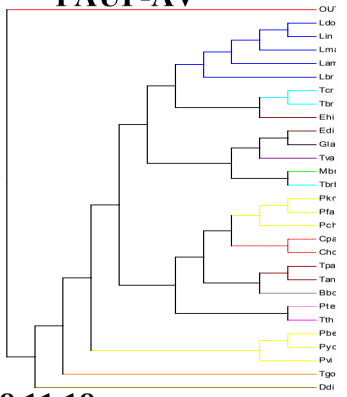
8.11.16

MRBAYES



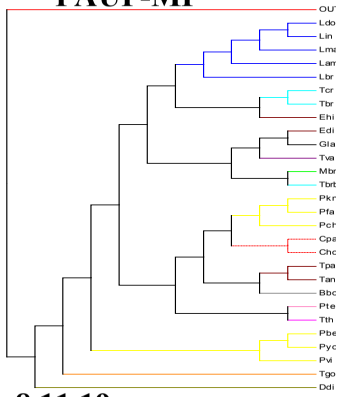
8.11.17

PAUP-AV



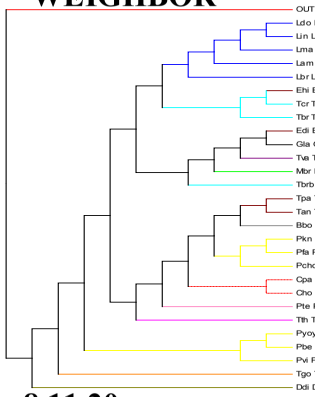
8.11.18

PAUP-MP



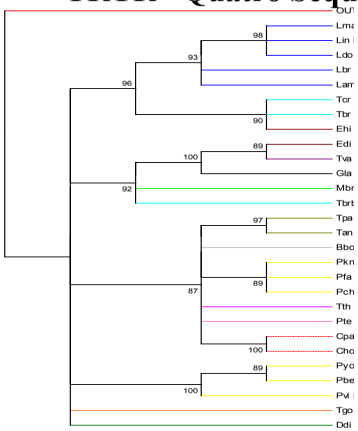
8.11.19

WEIGHBOR

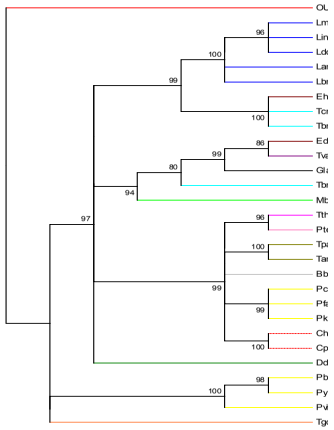


8.11.20

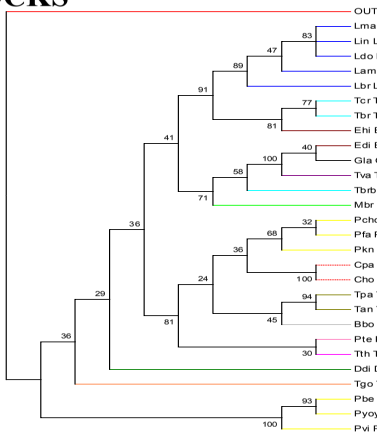
TRYR - Quatro Sequências Trimadas com o GBLOCKS



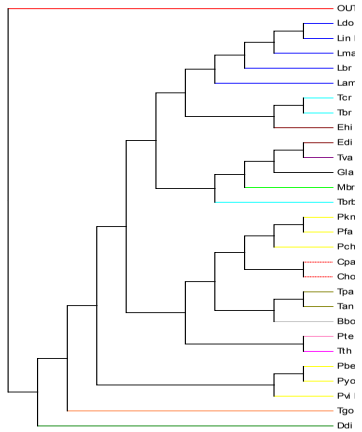
8.11.21



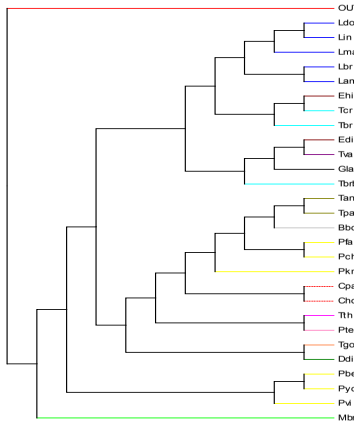
8.11.22



8.11.23

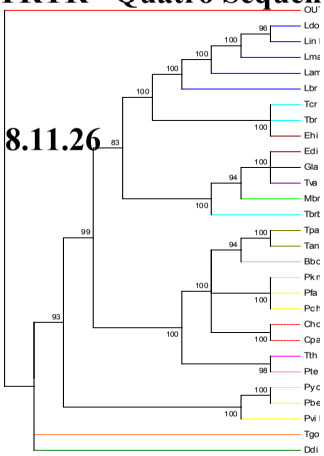


8.11.24

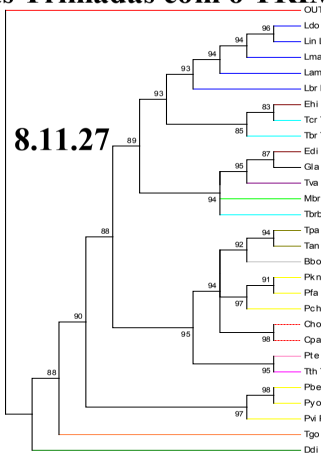


8.11.25

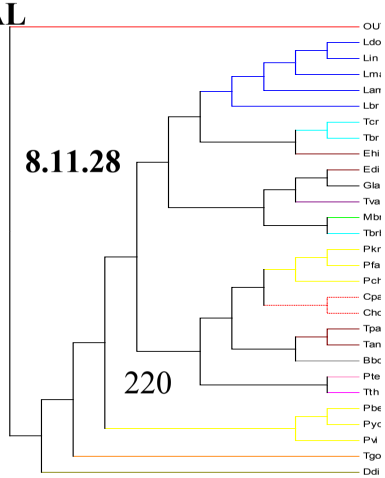
TRYR - Quatro Sequências Trimadas com o TRIMAL



8.11.26

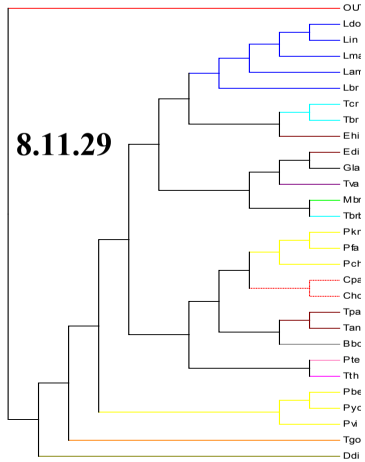


8.11.27

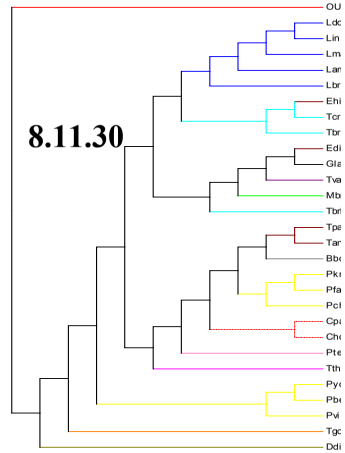


8.11.28

220



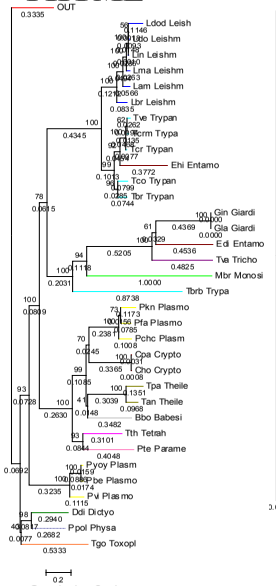
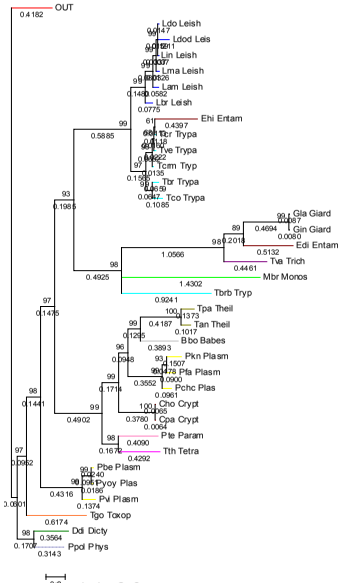
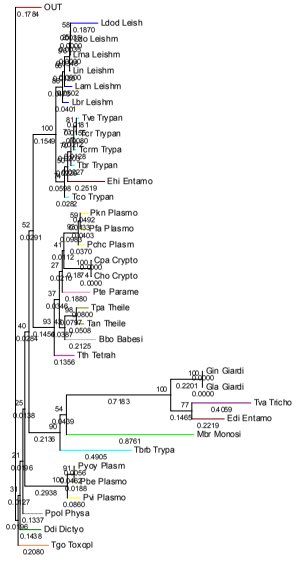
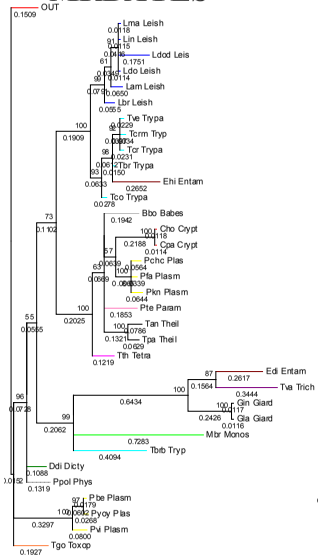
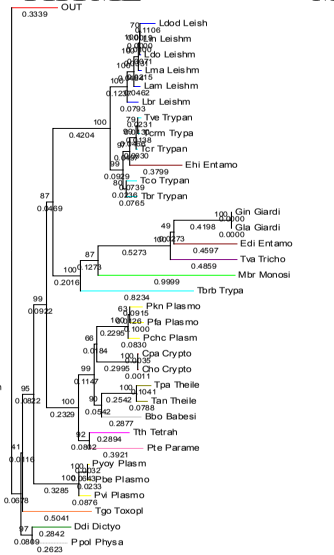
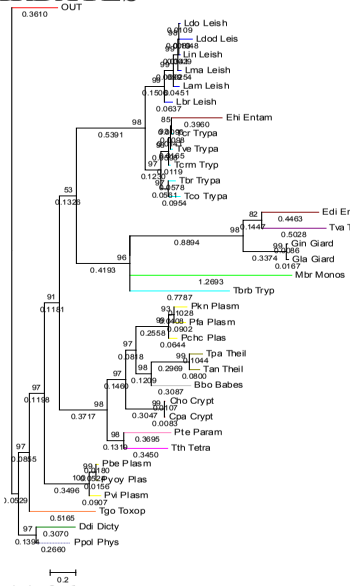
8.11.29



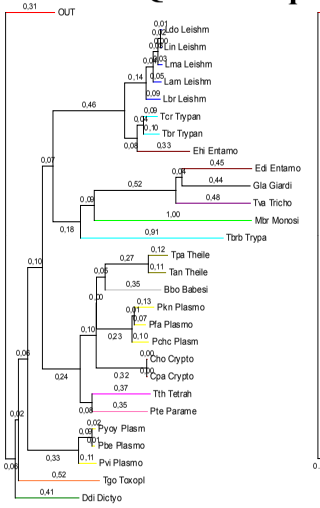
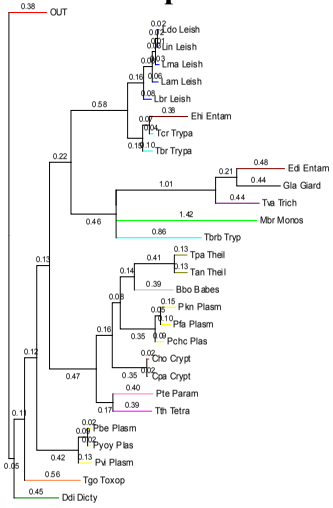
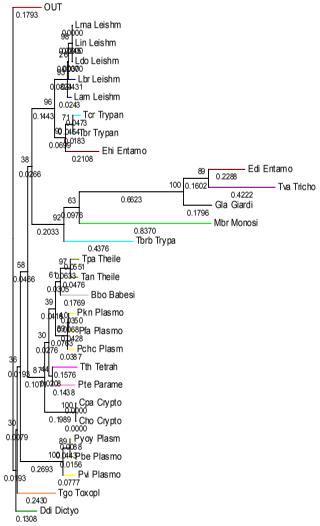
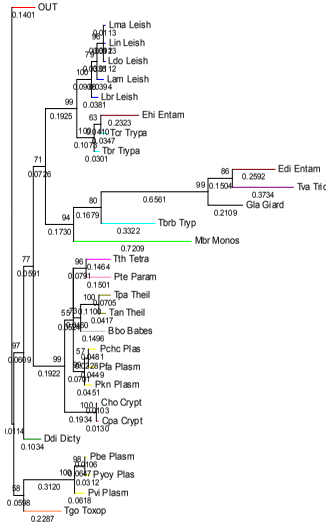
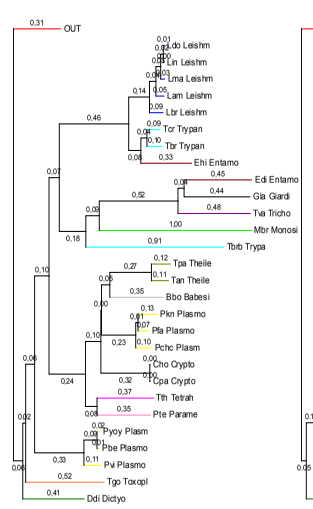
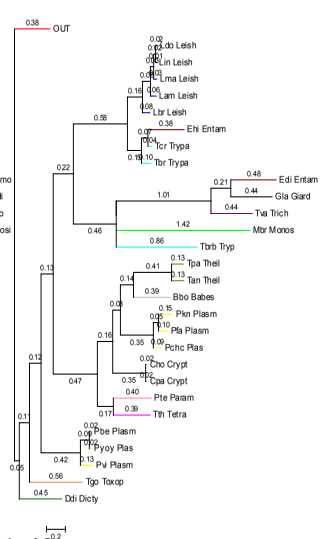
8.11.30

TRYR - Todas as Sequências Completas Trimadas com o GBLOCKS Trimadas com o TRIMAL

Mostrar só topologia

PHYML	MRBAYES	PHYML	MRBAYES	PHYML	MRBAYES
					
8.11.31	**8.11.32**	**8.11.33**	**8.11.34**	**8.11.35**	**8.11.36**

TRYR - Quatro Sequências Completas

					
8.11.37	**8.11.38**	**8.11.39**	**8.11.40**	**8.11.41**	**8.11.42**

Anexo 8.11 - TRYR - Árvores filogenéticas construídas com os programas PHYML, MRBAYES, PAUP-AV, PAUP-MP e WEIGHBOR

As árvores filogenéticas do gene TRYR foram construídas segundo a Metodologia M2 para os genes candidatos de resistência às drogas.

Todas as Sequências Completas:
- a *árvore* 8.11.1 representa o PHYML,
- a *árvore* 8.11.2 representa o MRBAYES,
- a *árvore* 8.11.3 representa o PAUP-AV,
- a *árvore* 8.11.4 representa o PAUP-MP,
- a *árvore* 8.11.5 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.11.31 representa o PHYML e
- a *árvore comprimento do ramo* 81.32 representa o MRBAYES.

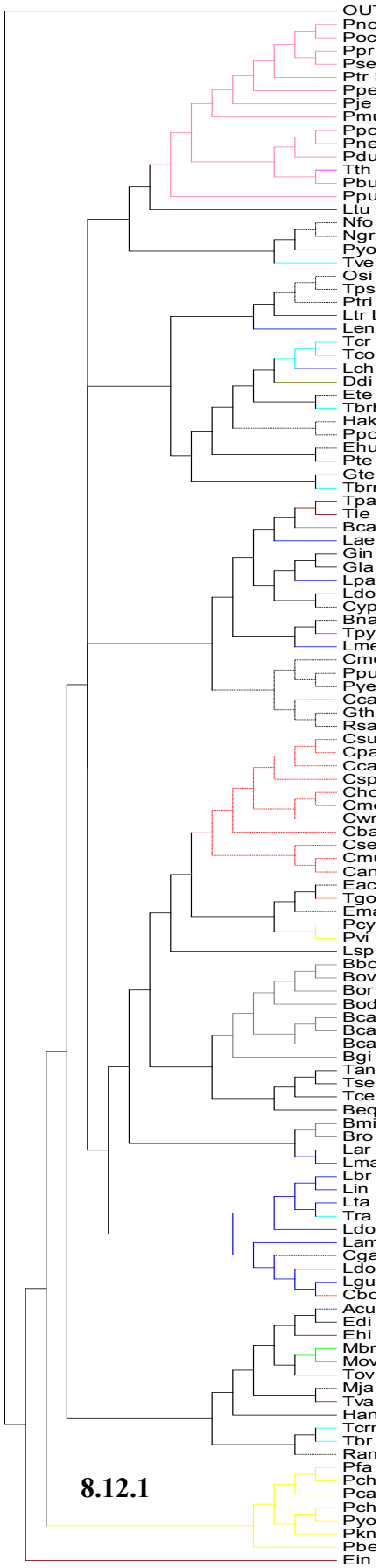
Todas as Sequências Trimadas com o GBLOCKS:
- a *árvore* 8.11.6 representa o PHYML,
- a *árvore* 8.11.7 representa o MRBAYES,
- a *árvore* 8.11.8 representa o PAUP-AV,
- a *árvore* 8.11.9 representa o PAUP-MP,
- a *árvore* 8.11.10 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.11.33 representa o PHYML e
- a *árvore comprimento do ramo* 8.11.34 representa o MRBAYES.

Todas as Sequências Trimadas com o TRIMAL:
- a *árvore* 8.11.11 representa o PHYML,
- a *árvore* 8.11.12 representa o MRBAYES,
- a *árvore* 8.11.13 representa o PAUP-AV,
- a *árvore* 8.11.14 representa o PAUP-MP,
- a *árvore* 8.11.15 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.11.35 representa o PHYML e
- a *árvore comprimento do ramo* 8.11.36 representa o MRBAYES.

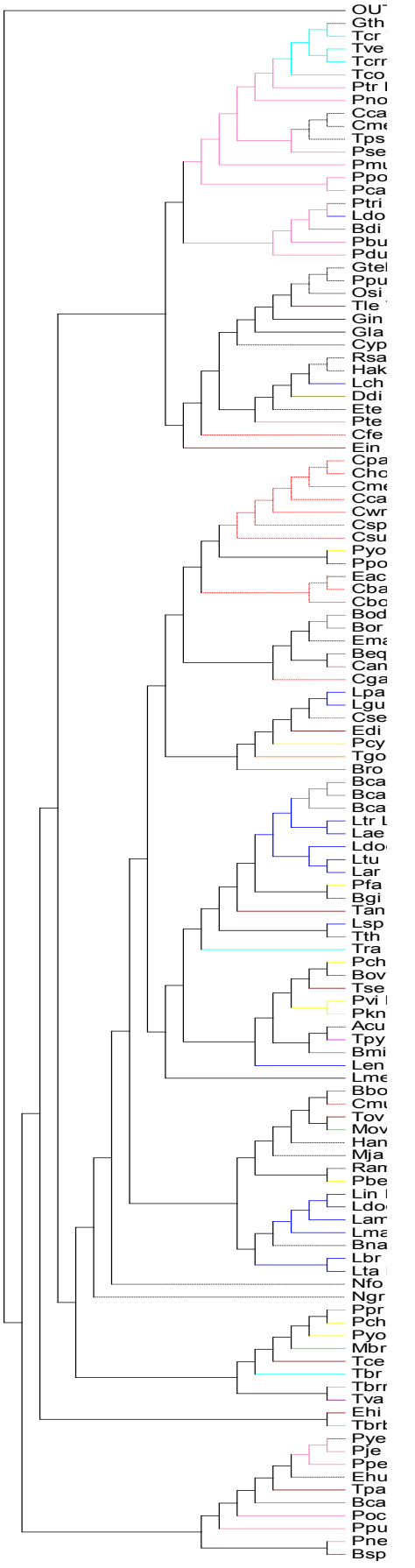
Quatro Sequências Completas:
- a *árvore* 8.11.16 representa o PHYML,
- a *árvore* 8.11.17 representa o MRBAYES,
- a *árvore* 8.11.18 representa o PAUP-AV,
- a *árvore* 8.11.19 representa o PAUP-MP,
- a *árvore* 8.11.20 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.11.37 representa o PHYML e
- a *árvore comprimento do ramo* 8.11.38 representa o MRBAYES.

Quatro Sequências Trimadas com o GBLOCKS:
- a *árvore* 8.11.21 representa o PHYML,
- a *árvore* 8.11.22 representa o MRBAYES,
- a *árvore* 8.11.23 representa o PAUP-AV,
- a *árvore* 8.11.24 representa o PAUP-MP,
- a *árvore* 8.11.25 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.11.39 representa o PHYML e
- a *árvore comprimento do ramo* 8.11.40 representa o MRBAYES.

Quatro Sequências Trimadas com o TRIMAL:
- a *árvore* 8.11.26 representa o PHYML,
- a *árvore* 8.11.27 representa o MRBAYES,
- a *árvore* 8.11.28 representa o PAUP-AV,
- a *árvore* 8.11.29 representa o PAUP-MP,
- a *árvore* 8.11.30 representa o WEIGHBOR,
- a *árvore comprimento do ramo* 8.11.41 representa o PHYML e
- a *árvore comprimento do ramo* 8.11.42 representa o MRBAYES.



8.12.1



8.12.2

Anexo 8.12 - A supermatriz e a superárvore

A árvore 8.12.1 representa a supermatriz (PAUP) e a 8.12.2 a superárvore (CLANN). Foram usadas Todas as Sequências Trimadas com o TRIMAL.

8.3 Filogenômica baseada na reconstrução da árvore de espécies de protozoários

Anexo 8.13-A - Genes ortólogos universais distribuídos em protozoários

COG	Anotação	Número de Genes
COG0012	Predicted GTPase, probable translation factor	131
COG0016	Phenylalanine-tRNA synthethase alpha subunit	58
COG0048	Ribosomal protein S12	153
COG0049	Ribosomal protein S7	81
COG0052	Ribosomal protein S2	87
COG0080	Ribosomal protein L11	117
COG0081	Ribosomal protein L1	94
COG0087	Ribosomal protein L3	117
COG0091	Ribosomal protein L22	56
COG0092	Ribosomal protein S3	62
COG0093	Ribosomal protein L14	144
COG0094	Ribosomal protein L5	85
COG0096	Ribosomal protein S8	106
COG0097	Ribosomal protein L6P/L9E	54
COG0098	Ribosomal protein S5	70
COG0099	Ribosomal protein S13	87
COG0100	Ribosomal protein S11	75
COG0102	Ribosomal protein L13	104
COG0103	Ribosomal protein S9	103
COG0172	Seryl-tRNA synthetase	53
COG0184	Ribosomal protein S15P/S13E	61
COG0186	Ribosomal protein S17	80
COG0197	Ribosomal protein L16/L10E	88
COG0200	Ribosomal protein L15	27
COG0201	Preprotein translocase subunit SecY	56
COG0202	DNA-directed RNA polymerase, alpha subunit	90
COG0256	Ribosomal protein L18	64
COG0495	Leucyl-tRNA synthetase	112
COG0522	Ribosomal protein S4 and related proteins	138
COG0525	Valyl-tRNA synthetase	53
COG0533	Metal-dependent proteases with chaperone activity	45

Anexo 8.13-B - Relação COG - KOG

Nome do COG/Descrição	Nome do KOG/Descrição
COG0012 / Predicted GTPase	KOG1491 / Predicted GTP-binding protein
COG0016 / Phenylalanyl-tRNA synthetase alpha subunit	KOG2783 KOG2472 KOG2783 KOG2784 / Phenylalanyl-tRNA synthetase
COG0018 / Arginyl-tRNA synthetase	KOG1195 KOG4426 / Arginyl-tRNA synthetase
COG0048 / Ribosomal protein S12	KOG1749 / 40S ribosomal protein S23/ KOG1750 / Mitochondrial/chloroplast ribosomal protein S12
COG0049 / Ribosomal protein S7	KOG3291 / Ribosomal protein S7
COG0052 / Ribosomal protein S2	KOG0830 / 40S ribosomal protein SA / KOG0832 / Mitochondrial/chloroplast ribosomal protein S2
COG0080 / Ribosomal protein L11	KOG0886 / 40S ribosomal protein S2 / KOG3257 Mitochondrial/chloroplast ribosomal protein L11/ KOG3504 / 60S ribosomal protein L29
COG0081 / Ribosomal protein L1	KOG1569 / 50S ribosomal protein L1 KOG1570 / 60S ribosomal protein L10A /KOG1685 / Uncharacterized conserved protein
COG0085 / DNA-directed RNA polymerase beta subunit/140 kD subunit	KOG0214 /RNA polymerase II, second largest subunit/ KOG0215 / RNA polymerase III, second largest subunit / KOG0216 / RNA polymerase I, second largest subunit
COG0087 / Ribosomal protein L3	KOG0746 / 60S ribosomal protein L3 and related proteins / KOG3141 / Mitochondrial/chloroplast ribosomal protein L3
COG0091 / Ribosomal protein L22	KOG1711 / Mitochondrial/chloroplast ribosomal protein L22/ KOG3353 / 60S ribosomal protein L22
COG0092 / Ribosomal protein S3	KOG3181 / 40S ribosomal protein S3
COG0093 / Ribosomal protein L14	KOG0901 / 60S ribosomal protein L14/L17/L23
COG0094 / Ribosomal protein L5	KOG0397 / 60S ribosomal protein L11 / KOG0398 / Mitochondrial/chloroplast ribosomal protein L5/L7
COG0096 / Ribosomal protein S8	KOG1754 / 40S ribosomal protein S15/S22
COG0097 / Ribosomal protein L6	KOG3254 / Mitochondrial/chloroplast ribosomal protein L6 / KOG3255 / 60S ribosomal protein L9
COG0098 / Ribosomal protein S5	KOG0877 / 40S ribosomal protein S2/30S ribosomal protein S5 / KOG2646 Ribosomal protein S5
COG0099 / Ribosomal protein S13	KOG3311 / Ribosomal protein S18
COG0100 / Ribosomal protein S11	KOG0407 / 40S ribosomal protein S14 / KOG0408 / Mitochondrial/chloroplast ribosomal protein S11
COG0102 / Ribosomal protein L13	KOG3203 / Mitochondrial/chloroplast ribosomal protein L13 / KOG3204 / 60S ribosomal protein L13a
COG0103 / Ribosomal protein S9	KOG1697 / Mitochondrial/chloroplast ribosomal protein S9 / KOG1753 40S ribosomal protein S16
COG0172 / Seryl-tRNA synthetase	KOG2509 / Seryl-tRNA synthetase
COG0184 / Ribosomal protein S15P/S13E	KOG0400 / 40S ribosomal protein S13 KOG2815 / Mitochondrial/choloroplast ribosomal protein S15
COG0186 / Ribosomal protein S17	KOG1728 / 40S ribosomal protein S11 KOG1740 / Predicted mitochondrial/chloroplast ribosomal protein S17
COG0197 / Ribosomal protein L16/L10E	KOG0857 / 60s ribosomal protein L10 KOG3422 / Mitochondrial ribosomal protein L16
COG0200 / Ribosomal protein L15	KOG0846 / Mitochondrial/chloroplast ribosomal protein L15/L10 KOG1742 / 60s ribosomal protein L15/L27
COG0201 / Preprotein translocase subunit SecYa	KOG1373 / Transport protein Sec61, alpha subunit
COG0202 / DNA-directed RNA polymerase alpha subunit/40 kD	KOG1521 / RNA polymerase I and III, subunit RPA40/RPC40 / KOG1522 / RNA polymerase II, subunit POLR2C/RPB3
COG0256 / Ribosomal protein L18	KOG0875 / 60S ribosomal protein L5 KOG3333 / Mitochondrial/chloroplast ribosomal protein L18
COG0495 / Leucyl-tRNA synthetase	KOG0435 KOG0437 / Leucyl-tRNA synthetase
COG0522 / Ribosomal protein S4 and related proteins	KOG3301 / Ribosomal protein S4 KOG4655 / U3 small nucleolar ribonucleoprotein (snoRNP) component
COG0533 / Putative GTPases (G3E family)	KOG2707 KOG2708 / Predicted metalloprotease with chaperone activity (RNAse H/HSP70 fold)

Anexo 8.13-C - Distribuição dos genes ortólogos universais nos genomas de protozoários

Espécies	COG 0012	COG 0016	COG 0048	COG 0049	COG 0052	COG 0080	COG 0081	COG 0087	COG 0091	COG 0092	COG 0093	COG 0094	COG 0096	COG 0097	COG 0098	COG 0099	COG 0100	COG 0102	COG 0103	COG 0172	COG 0184	COG 0186	COG 0197	COG 0200	COG 0201	COG 0202	COG 0256	COG 0495	COG 0522	COG 0525	COG 0533
Aca	-	-	x	-	x	-	-	-	-	-	x	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Apo	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ahe	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bbo	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Bna	-	-	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	-	-
Cro	-	-	x	-	-	-	-	-	-	-	x	-	x	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cri	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Csy	-	-	x	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cho	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x
Cpa	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x
Cme	-	-	x	x	x	-	x	x	-	x	x	x	x	x	x	x	x	x	x	-	-	x	x	-	x	-	-	-	x	-	-
Cca	-	-	x	x	x	x	x	x	-	-	x	x	x	x	x	x	x	x	x	-	-	x	x	-	-	-	-	-	x	-	-
Cpar	-	-	x	x	-	x	x	x	-	x	x	x	x	x	x	x	x	-	x	-	-	x	x	-	-	-	-	-	x	-	-
Dvi	-	-	x	-	-	-	-	-	-	-	x	-	x	x	-	x	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-
Dci	-	-	x	-	x	x	-	-	-	-	x	-	-	x	-	-	-	-	-	-	-	-	x	-	-	-	-	-	x	-	-
Ddi	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Dfa	-	-	x	-	x	x	-	-	-	-	x	-	-	x	-	-	-	-	-	-	-	-	x	-	-	-	-	-	x	-	-
Ddic	x	x	x	-	-	-	-	-	-	-	x	-	x	x	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ete	-	-	x	-	-	x	-	-	-	-	x	-	-	-	-	-	-	-	-	x	-	-	-	-	-	-	x	-	x	-	-
Ehu	-	-	x	x	x	-	-	x	-	x	x	x	x	x	x	x	x	-	x	-	-	x	x	-	x	-	-	-	x	-	-
Ehi	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Edi	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x
Egr	-	-	x	x	x	-	-	-	-	x	x	x	x	-	-	-	x	-	x	-	-	-	-	-	-	-	-	-	x	-	-
Elo	-	-	x	-	x	-	-	-	-	-	x	x	x	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-
Fve	-	-	x	-	-	-	-	-	-	-	x	-	x	x	-	x	-	-	-	-	-	-	-	-	-	-	-	-	x	-	-
Gla	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x
Gth	-	-	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	-	-	-
Gte	-	-	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	-	-	x	x	-	x	-	-	-	x	-	-
Han	-	-	x	x	x	-	x	-	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	-	x	x	-	x	-	-
Hak	-	-	x	x	x	x	x	x	-	-	x	x	x	x	x	x	x	x	x	-	-	x	x	-	-	-	-	-	x	-	-
Ldi	-	-	x	-	-	-	-	-	-	-	x	-	x	x	-	x	-	-	-	-	-	-	x	-	-	-	-	-	x	-	-
Lve	-	-	-	-	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Lbr	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Ldo	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	-	x	-	-	-
Len	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-
Lin	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Lma	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x

Mja	-	-	x	x	x	x	-	-	-	-	x	x	x	x	-	x	-	-	-	-	-	-	-	-	-	-	-	-	x	-	-
Mbr	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Ngr	-	-	x	-	-	-	-	-	-	-	x	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Oda	-	-	x	-	-	-	-	-	-	-	x	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Osi	-	-	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	-	-	x	x	-	x	-	-	-	x	-	-
Pau	-	-	x	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pca	-	-	-	-	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Pte	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Ptr	-	-	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	-	-	x	x	-	x	-	-	-	x	-	-
Pin	-	-	x	-	x	-	-	-	-	-	x	-	x	x	-	x	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-
Pra	-	-	x	-	-	-	-	-	-	-	x	-	x	x	-	x	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-
Pso	-	-	x	-	-	-	-	-	-	-	x	-	x	x	-	x	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-
Pbe	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pch	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x
Pfa	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pkn	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pvi	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x
Pyo	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x
Ppa	-	-	x	-	x	x	-	-	-	-	x	-	-	x	-	-	-	-	-	-	-	-	x	-	-	-	-	-	x	-	-
Ppu	-	-	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	-	-	x	x	-	x	-	-	-	x	-	-
Pye	-	-	x	-	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	-	-	x	x	-	x	-	-	-	x	-	-
Pli	-	-	x	-	x	-	-	-	-	-	x	x	x	x	-	x	-	-	-	-	-	-	-	-	-	-	-	-	x	-	-
Ram	-	-	x	x	x	x	x	-	-	-	x	x	x	x	-	x	x	-	-	-	-	-	-	-	-	-	-	-	x	-	-
Rsa	-	-	x	-	x	-	x	x	-	x	x	x	x	x	x	x	x	x	x	-	-	x	x	-	x	-	-	-	x	-	-
Sfe	-	-	x	-	-	-	-	-	-	-	x	-	x	x	-	x	-	-	-	-	-	-	x	-	-	-	-	-	x	-	-
Tth	x	x	x	x	x	x	x	x	x	-	x	-	x	x	-	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x
Tma	-	-	x	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tpar	-	-	x	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tpi	-	-	x	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tpy	-	-	x	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tps	-	-	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	-	-	x	x	-	x	-	-	-	x	-	-
Tan	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Tpa	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Tgo	-	-	x	-	x	-	-	-	-	-	x	-	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	x	x	-	-
Tva	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Tbr	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Tcr	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Anexo 8.14 - Número de GI das sequências usadas nas filogenias

COG0012

62359197 62359712 134063269 134063323 134065023 134065959 134066211 134066282 134070784 134070825 134072371 134073687 134074076 134074188 7658164 23504826 194247147 193809639 194247306 32398750 70800090 70800143 70906111 74831277 74831283 74831330 74831349 37729656 67591456 66815061 66827157 66810259 66805123 66801213 67470854 67484586 159110528 159111683 159115573 73544489 73544595 157874309 72550046 157876922 157877092 68073569 68067551 68070105 70948715 124506405 124511984 124513898 156093524 156095358 156092014 156091836 83286382 82539543 82705616 118364117 118376292 118382824 146183354 71032329 71032369 71027925 74026362 84043672 84043784 71745170 71747732 71747846 156083024 156083691 156085942 156085980 156087895 67593816 67624197 67599391 66362270 66356360 66359460 66475434 167391739 167395011 154339988 154340096 154343491 154345361 154345864 154346006 146091024 146091169 146097008 146102251 146104007 146104415 167527311 167530586 167520332 167522160 167522751 167523461 167523900 145537167 145553341 145527963 145529075 145529411 145534418 145534714 145489801 145498321 84999412 84999454 85001163 123415144 123464697 123438916 71402860 71661506 71400644 71401680 71654923 71649554 71414858 71419936 71395415 71407364 71659517

COG0016

28829368 134064785 134072138 23496841 23510671 193807542 194247649 193811544 78190821 50057625 46362269 66807259 66820560 67472262 159107426 157873803 68071149 68070951 70950809 70929092 70949951 70924242 124505829 86170496 124806408 156094999 156100737 156095643 82540861 82539580 82541749 118372908 118389432 71030178 71029228 74025804 156085651 156088893 67623789 67624611 66359410 126652720 167393910 154343015 146096330 167533149 167536178 124088790 145507648 145544557 145474015 145478091 84995656 84997085 154414980 154414982 154414986 71664621

COG0048

562049 110810221 1040985 160688798 1019631 578752 1171607 160688749 114150054 114150100 114150144 10839 5231249 4958894 28829070 134061872 134061873 134071603 134071604 10178685 23344065 10444216 4493905 6562755 194247082 193809045 2258332 5306138 15011475 75756360 32398723 187479919 1088425 31322469 28786967 50057404 51339682 11467068 115443576 161899279 11466309 9653256 11466491 8954390 30468207 11465447 11467352 84508599 7524991 66823621 187764090 84508520 71842335 41203481 67466775 67465599 67463378 67483053 11466995 12545423 84508559 159109834 51209918 11467742 162606492 189095409 21449994 157869012 157869014 11466647 11466181 11465895 11467445 8928601 118411084 9695388 145932437 145932344 68076897 68071153 70954208 70916985 70948927 124505381 124504805 156094529 156094107 82596441 82595242 60117118 11465755 11465645 90994479 15150724 11466502 149072034 11466598 50261281 50261309 114329821 114329867 114329911 11466262 15027639 146163441 118411216 74325182 71032767 71026388 71025960 71746622 71746624 156082334 156086446 156087651 126649277 31442377 167384968 167390295 167395258 167385801 160331444 186920138 154337202 154337204 146094256 146094258 167533598 23464602 124088230 145504390 145505029 145505714 145507226 145545007 145474545 145487884 145496579 84999086 85001435 11496546 123501267 123463800 123488090 123478863 123478903 154418783 154419638 154423005 71403979 71403981 71661489 71661499

COG0049

110810220 2350900 134060018 134060019 134067984 134067985 10178682 4039122 194247140 2258333 75756504 32399008 28565604 28565605 115443575 161899567 67591273 126165893 11465448 11467351 66808411 71842336 67473874 67477084 67468810 67484034 11466994 159107490 51209917 11467743 162606064 189095408 157865716 157865718 11466644 11467444 118411085 68003826 68067854 70945908 124511970 156093534 82704680 11465754 90994478 11466503 149072035 118353446 118411217 71026645 71755579 156082401 66475896 167377621 167383907 167392227 160331825 154333504 154333506 146079894 146079896 167523581 145501238 145501715 145511227 145529303 145530634 145534506 145495306 145498361 85001648 123485933 123421567 154417175 154417181 154417227 154417229 71659005 71659007 71666482 71666484

COG0052

562040 110810233 160688816 160688767 4958912 2350896 2350932 134066495 134074263 134074273 10178696 23495130 193808246 2258384 22035888 4809047 34305123 75756486 187479910 51339700 11467059 115443588 161899531 30468231 11465574 11467370 7525009 66826691 187764108 71842276 67462433 67483019 11467017 12545419 159114778 51209981 11467687 162606370 189095356 157877540 157877562 11466658 11467570 118411020 9695410 68069781 70948671 124802649 156082017 82704653 60117136 11465715 90994440 11466554 149072059 146165118 118411126 71029482 74025154 74025178 156089135 67624811 66359200 167377034 167395276 160331578 186920126 154346432 146104702 146104747 167515424 167515608 145537035 145549468 145553453 145510154 84997323 123457462 123427072 123439029 123475378 154412101 71403370 71406414 71406749 71396821 71402795

COG0080

160688796 160688747 4958892 2350869 62359726 62359736 62175142 134062541 134063311 134065658 134069195 134070122 134070813 10178665 23496037 23504656 193809177 193809718 2258327 32307496 75756392 187479921 1088423 31322473 70800131 70905804 51339680 161899343 30468130 11465493 11467314 7524989 66806151 187764088 67465703 67473968 67464170 159116773 51209889 11467611 162606074 189095318 157870377 73544571 72547997 11466627 11467549 118411002 68074081 68075663 68008782 68005679 70949841 70953539 70914242 124506259 124803673 156098390 156095506 82540651 83273774 60117116 11465803 90994528 11466497 149071969 118400634 146161229 118411123 71033335 71028548 84043734 84043754 72393265 71745992 156086990 156084702 67616823 66359054 87043011 126022809 167393038 167376963 167393873 160331490 154338535 154340072 154344759 146084778 146088222 146091124 167525156 145502621 145523445 145524052 145525008 145511612 145476333 145485929 145535291 84996459 84998532 123509555 123495458 123408132 154415055 154419236 154420621 123453894 71417151 71661532 71414154 71661043 71661069 71402841 71650641 71413453

COG0081

1123023 28828132 2350966 42491237 134061025 134066361 134068974 134074009 23497461 23496632 193807417 193810534 193810680 2258328 4193363 32307584 75756381 161899321 30468131 11465492 11467315 66818615 67476939 183230142 159114702 51209890 11467612 162605892 189095317 157867831 157877260 11467548 118411001 68076911 68065089 70952237 70953609 70953658 124805683 124809402 124511778 156100503 156101758 156094882 82595793 82596540 82595251 68525556 11465804 90994529 11466498 149071970 146184970 146162404 118411124 71033067 71033997 71028234 74024938 71748990 156086964 156087068 156087404 156088087 67621484 67623427 66363094 66360018 167539866 167539880 160331217 154335511 154346164 146083933 146103761 167525004 167523567 145545518 145510009 145490193 145495248 145498303 84996167 84997862 84998510 84998818 123445283 123402053 123487009 154419768 123975451 71660333 71654412 71420692

COG0087

167816 62359480 62359481 62360659 134061564 134063675 134065024 134065025 134070872 134072372 134072373 134073003 23495138 23496969 23505082 193808256 193808523 193811677 3861468 32307490 75756351 66476128 70799538 78190797 161899261 30468182 11465422 11467343 66805433 66802278 71842315 67475438 67472477 183235559 67473369 159109947 51209943 11467717 162605876 189095434 73536476 157874311 157876054 157876056 157876058 157876060 126022788 118411058 68066851 68073103 68065494 68011480 70943067 70924866 70927424 70951495 124506912 124802670 124806820 156081997 156097364 156095889 82753677 82596814 82593828 11465780 90994504 149072009 146185729 118389862 118411190 71032875 71027797 71029452 72387453 72387810 72387812 156082902 156083276 156089107 67594775 126649185 167386238 167377647 167388407 160331207 154336587 154340798 154343493 154343495 146092753 146097010 146097013 146099403 167535326 167518906 145525954 145527660 145475213 145517081 145532966 145497857 84997293 84998982 84998984 85000981 123440132 123445156 123495590 154419020 154422961 123470017 71652016 71652018 71408950 71416007 71411062

COG0091

134062321 134064740 134069969 134072092 193809908 66809169 67473048 67473573 67472535 159114764 162605890 157869943 157873703 68062668 68074749 70951972 124513742 156096516 82753814 146170683 71031290 71755285 71749196 156083152 67624507 66360360 167376128 167378219 167395392 160331219 154338095 154342925 146087652 146096179 167515572 145500438 145509905 145526865 145527142 145527606 145514255 145514562 145475255 145533006 145494812 145497343 145497791 84994548 123446833 154412855 123478845 123468434 71651335 71421846 71404313 71666237

COG0092

1688320 134060568 134065214 134068526 134072541 23497701 8247324 193810280 75756337 78190807 161899235 30468188 11467339 66800843 71842320 67467661 67484576 11467009 159117007 51209937 11467723 162605846 157866908 157874873 12751545 118411065 68010962 68070737 70947163 124810210 156100003 11465774 90994498 149072015 146184763 118411197 71028826 71744718 71748612 156084414 67614912 66362866 167394938 167389197 160331719 154334600 154343872 146082044 146097654 167519583 145541820 145536261 84996723 123456864 123444699 154413076 123476890 154415577 123483338 71408264 71653987 71656239

COG0093
562055 552049 110810196 160688805 13285 1171600 160688756 114150086 114150131 114150176 7321603 13857 5231245 4958901 62360509 134061657 134065817 134065818 3851618 134073087 134073556 10178666 23344076 10444236 23504678 193809694 193810093 2258344 5306169 15011507 75756416 187479915 2627253 31322463 51172914 70905963 70905964 51339689 1536839 11467074 115443551 161899391 11466306 11466490 8954384 30468192 11465432 11467336 84508602 7524998 66803967 187764097 84508523 71842323 67470668 67474853 67472184 11467012 12545440 84508562 159109758 51209933 11467727 162606006 189095424 21449998 72549131 72549135 157876257 11466628 11466201 11465894 11467460 8928583 118411069 9695373 145932424 145932330 68068641 68075865 70930748 70945029 124506303 124513388 156096144 156095462 82704672 82794048 60117125 11465770 90994494 15150727 11466514 149072019 11466570 50261285 114329853 114329898 114329943 11466293 15027671 146164168 118411201 74325184 71030494 71028768 72387638 71745554 71745556 156082346 156084630 156088723 67594829 66475246 31442371 167384703 167380314 167385974 160331023 186920143 154336773 154345077 154345079 146099698 146101722 167534168 23464613 145527302 145475561 145533336 145497491 84995348 84996671 11496542 123352746 123455292 123455294 123455296 123509562 123487174 71659840 71659842 71663945 71395987

COG0094
110810197 295738 559581 577833 1333641 134059056 134061962 134066965 134069622 12311882 10178668 194247120 2258345 75756315 32398798 78190795 1246369 115443552 161899191 30468194 11465434 11467335 66815425 71842324 67478913 67483142 67471014 67478957 11467013 12545415 159113140 51209931 11467729 162606144 189095422 76363693 157869211 11466630 126022787 118411071 68075317 70943404 124511928 156093566 83314399 11465768 90994492 15150728 11466515 149072021 118360820 118411203 71032475 71744960 71744962 156086132 67619305 66475518 167392809 167381101 167375730 167386027 160331147 186920144 154331838 154337381 146076266 146086348 167516410 145499984 145538824 145528562 145475927 145488693 145495097 145498891 145499343 84999322 123478647 123427769 123976993 123379265 123481292 123478990 71401066

COG0096
562058 110810217 881426 28850259 42491241 42491231 62358889 134060041 134061044 134064349 134068005 134068990 134071070 10178673 193808961 2258347 75756280 70799722 3758866 11467077 115443572 161899121 11466321 126165894 11465435 11467334 84508593 66818863 84508514 71842325 67468666 67476737 67476841 11467014 12545416 84508553 159110821 51209930 11467730 162606566 189095421 21449988 157865756 157867865 73536854 11466635 11465887 11467457 118411072 9695408 145932461 145932371 68065258 68071429 70925667 70947548 124504993 156097021 82594213 82595101 11465767 90994491 15150715 11466517 149072022 50261297 118360340 118411204 74325177 71029982 71755547 72387207 71748952 156089589 67594923 66361601 167379180 167391992 167376877 160331369 186920117 154333550 154335549 154342144 146079973 146083999 146093191 167520234 145537588 145540900 145524397 145549265 145528057 145516050 145483251 145488111 145491686 145519239 84995840 123490281 154413044 123482400 71401957 71409235 71664580 71666518

COG0097
160688809 160688760 4958905 62358789 4590398 134061871 134064287 134071551 134071605 10178700 193810183 2258348 75756322 84682137 2627257 51339693 161899205 11466322 67591332 30468196 11465436 11467333 84508594 7525002 66828929 187764101 84508515 71842326 67473391 67466441 84508554 159111130 51209929 11467731 162606384 189095420 21449989 157869010 157872833 11466662 11467456 118411073 9695409 145932462 145932372 68072635 70953061 124513230 156099270 83033193 60117129 11465766 90994490 15150716 11466518 149072023 11466589 50261298 118395507 118411205 74325178 71032597 72390583 71746626 156086254 66356646 87043037 167391497 167377627 167388321 160331269 160331592 186920118 154337200 154342021 146094152 146094260 167538171 145500002 145528682 145475955 145488673 145495003 145499367 84999220 123421312 123337227 123506838 154418851 123976818 71403977 71649657 71661487 71407652

COG0098
7299 2350878 2350880 2350948 2746717 134061167 134064742 134070900 134072094 18699610 23497519 193810477 32307554 75756482 78190805 161899523 67591835 30468198 11465438 11467331 66800083 71842327 67465591 159111522 51209927 11467733 162605774 189095418 157868060 157873707 11467454 118411075 68067285 70941721 124809606 156100389 82752715 11465764 90994488 149072025 146169362 118411207 71031410 71749200 71749220 156084069 66356488 167390301 167382049 160331426 154335794 154342929 146092851 146096186 167523024 145503960 145506206 145539163 145542588 145480933 84994434 123412904 123480539 123444423 123473854 154418683 71651337 71651351 71406483 71421842

COG0099
509603 28829919 134066069 134066070 134073879 134073880 10178675 10444242 23496197 193809336 2258351 75756417 66476122 161899393 30468201 11465441 11467401 84508609 66818921 84508531 71842330 183231655 67478589 183233172 84508569 159107620 51209924 11467736 189095415 21450006 157876640 157876642 11466637 11466207 12751544 118411078 9695400 145932453 145932363 68069309 70938520 70943818 124804238 156098723 82794396 11465761 90994485 15150734 11466521 149072028 50261289 118372788 118411210 74325175 71032473 71747446 71747448 156086130 67616771 66359256 167378734 167381370 167379766 186920119 154345580 154345582 146103213 146103218 167527167 145539402 145549051 145512882 145479337 145483041 145534229 84999324 123492382 123507646 123479230 123448087 123367594 123482954 123503102 123475112 123453508 71666187 71666149

COG0100
110810215 162240 2350992 42491239 42491229 62359856 134063453 134064310 134084929 134071583 23504648 193809726 2258352 75756353 115443570 161899265 30468202 11465442 11467400 66802868 71842331 67475755 183230114 11467041 159111843 51209923 11467737 162605868 189095414 157871656 157872891 11467450 118411079 68072747 70938201 124506243 156095522 83273504 11465760 90994484 11466522 149072029 146182837 118411211 71028538 72390641 156084712 67623403 66362484 167393430 160331197 186920120 154340355 154342067 146091837 146094216 167519138 145502717 145523347 145523966 145506869 145524914 145511411 145476241 145485845 145517969 84996447 123455508 123487441 123489921 123472295 123410181 71400062 71415861 71664001

COG0102
28850421 28828553 23503402 62358756 62360518 62176316 2612931 134060491 134061404 134061642 134068452 134072830 134073072 23494906 193807723 193808739 32307500 75756250 161899061 30468204 11465444 66821519 66819767 67470612 67480065 67469617 183234108 67467212 159109125 51209921 11467739 162605982 189095412 157866762 157875585 157876227 11467448 118411081 68076449 67983599 68070607 70932631 70919057 70951318 70947086 70914992 124801264 124801997 156093731 156096629 82793887 11465758 90994482 149072031 118377465 146181945 118411213 71032509 71028758 72389080 72387666 72388162 156086164 156086180 156084640 67624381 66357740 167391032 167540042 167384126 167385776 160330997 154334446 154336267 154336743 146081814 146098758 146099647 167535340 167515562 145503966 145542582 145528472 145477345 145480925 145486164 145488354 145490620 145495183 84996661 84999290 123411974 123504865 123505737 123479117 154416881 123494502 154422729 71410787 71416440 71424291 71649293 71657094 71407626

COG0103
62176151 62176152 134062904 134062905 134070462 23497197 23496307 193809448 193810937 193811299 3746417 3746423 3695262 95007312 32307508 75756366 28787567 66476132 161899291 30468205 11465445 11467398 66810045 66800973 71842333 67480365 183231489 67483152 67466982 11467026 12545422 159112547 51209920 11467740 162605934 189095411 157871085 157871087 11467447 118411082 68072891 68072945 68070051 70950873 70951553 70948905 124804601 124808408 124512506 156102252 156098941 156102971 82915031 82793272 82705805 11465757 90994481 149072032 146163186 146183499 118411214 71032171 71030542 72390868 72390870 156086112 156087611 156088675 67608989 66358632 167381093 167393081 167378354 167385853 160331203 154339259 154339261 146089618 167525180 167533983 145523495 145510060 145527244 145529275 145475615 145533384 145534538 145485989 145497429 84995300 84999594 85001488 123410600 123418393 123468527 154415419 154421138 154421152 154422037 71418495 71418497 71424424 71424421

COG0172
133950600 28828372 42733938 109238514 134061152 134067901 23496687 194247104 193811385 75756430 161899419 67604545 66827607 66823277 66821029 183231706 183233620 67469401 159110869 159119702 162606512 157865542 68073951 68063117 70941484 124805876 124511894 156094491 156103129 83317733 146163080 71032093 71755749 156083631 156086478 66361666 118490844 167385191 167390181 167392442 160331655 154335764 146079628 167534977 167524046 145509547 145494412 84999662 85000409 123417088 123463833 71407799 71421977

COG0184
1401205 110681630 62358549 134061201 134065440 134067618 134069098 134072742 194247307 75756454 161899451 66809663 66804969 183231337 67481543

159114668 162606334 157865072 157868126 157875365 68073571 124513900 156101355 82539545 82593966 146185324 71032761 84043888 71749292 156083026 67615320 66357962 167381855 167395475 160331293 154335862 154344324 146078395 146084420 146098263 167515792 145524409 145548060 145528087 145516058 145488101 145491696 145519269 145497019 84999094 123483596 123479951 123975177 123505067 123470781 123486522 154421682 71408024 71421892 71403504 71417376

COG0186

13177633 28828614 134061908 134065823 134069383 134069550 134073561 5764667 193808953 193809014 83642498 75756384 70905969 7768291 161899327 30468191 11465431 11467337 66811940 66818058 71842322 67480803 67466589 159108538 51209934 11467726 162605740 189095425 157868717 157869112 72549163 11467461 118411068 68067397 70937128 124505009 156097005 82752681 11465771 90994495 149072018 118395679 118400616 118411200 71030444 115504447 71745544 71746518 156084994 156089415 167376448 167390487 160331051 154337274 154345089 146085475 146086110 146101737 167525306 145524361 145547360 145548156 145528013 145516012 145488023I 145491782 145519199 145496929 84995402 123394395 123425501 154412006 154415799 123472884 123408443 71659828 71415083 71403962 71414429 71411518

COG0197

110810195 160688804 160688755 4958900 134059077 134059098 8977991 134066994 134067015 12311801 12311821 23497206 193810923 14091455 75756377 32398995 187479916 2627252 21309663 78190793 51339688 115443550 161899313 67592553 30468189 11465429 11467338 84508605 7524997 166240289 187764096 71842321 183237136 67472614 67473511 159111635 51209936 11467724 162606496 189095427 21450002 76363750 76363790 11467463 118411066 9695404 145932457 145932367 68066022 70950044 124808442 156102234 82794128 60117124 11465773 90994497 149072016 50261293 118358504 118411198 71030040 71745040 71745090 156089531 66475872 167393484 167379745 167383951 160331448 154331880 154331922 146076384 146076468 167517447 145539394 145512866 145479333 145534243 84995786 123449164 123449205 123487056 154417689 123411254 71422972 71417265 71653448 71419116

COG0200

161842 161844 2232029 62176751 4104078 13276187 13276189 13276205 13276207 134060629 134065814 134065816 134068584 134073553 134073555 23497338 23497344 194247498 193810800 193810806 32307618 75756375 790954 70905960 70905962 46361134 161899309 66801351 66809775 67469831 183232475 183233947 67478373 159107543 162606510 157867049 72549120 72549128 68076535 68074795 68070907 70946477 70953398 86171217 124808927 124808949 156101968 156101982 156101005 82915191 82596519 83317937 118373022 146168461 71033131 71030482 71026773 72389546 71745558 71745562 156087112 156088233 156088757 67598841 66359520 167382469 167385411 167390283 167387447 160331653 154334721 154345071 154345075 146082364 146101711 146101718 167516096 167519024 145537229 145541193 145549666 145553273 145510338 145530151 145486513 145489566 145494069 145520142 84995360 84998752 85000031 123479634 123469153 123396945 154415797 154418905 123382236 71419650 71659844 71659846 71408480

COG0201

134060027 134067993 3057044 193809966 52352493 67591486 30468199 66828895 71842328 183232362 159110159 51209926 11467734 157865734 11467453 118411076 67983419 68069049 70949576 70926576 124513638 156096412 83314515 11465763 90994487 149072026 118400925 146161267 118411208 71026625 71755565 156082373 66362634 167384871 167381568 167383640 154333522 146079928 167534051 145502234 145522963 145545494 145526322 145511658 145516705 145535227 145489823 145498704 85001629 123459759 154416678 154417534 123494221 123410308 71658989 71666500

COG0202

2654270 62360648 21842208 134061228 134063686 134069124 134070883 23497770 23496369 23505130 193808570 193809507 193810209 193811119 88687066 75756371 75756372 70799549 161899301 161899303 66812176 66801649 67471071 159115523 159116466 162605758 162606468 157868180 73536498 68076655 68069197 68075539 68000767 68062152 68070451 70945266 70943148 70945964 124507008 124804792 124810391 124512894 156099322 156097458 156099087 156102625 83282536 83286741 83314615 82540505 146164033 146169387 71031829 71032023 71031382 71027147 72387431 71749338 156082852 156083567 156089709 156089735 67611563 67624207 66362260 66356630 167386476 167387776 160331409 160331749 154335916 154340820 146084517 146092793 167535426 167517689 145506343 145528814 145531944 145517526 145518624 84994468 84999732 84999910 85000353 123448560 123464689 71656797 71423439 71663692

COG0256

2350902 2350904 42491233 42491225 134065626 134065627 134073376 23497297 194247543 193810846 75756516 51172912 70905771 70905772 70905773 78190799 46361086 75875724 161899591 66816906 67478072 67469033 159112008 162606588 72547732 72547735 72547739 68065284 68070119 70950764 70948198 86170968 124808771 156102056 156100913 83273592 82705794 118356685 71030018 71746208 71746210 156089549 67597151 66356654 167383135 167384940 160331305 154344695 154344697 146101020 167519038 145506266 145546360 145515701 145487326 84995804 123508736 154411960 123975295 123427377 123473909 71652057 71406222 71406224

COG0495

161801 28828225 57157191 3219290 57157195 57157213 62358810 15487676 134060244 134064262 134066545 18021914 134068219 134071527 134074315 23497663 193807873 194247454 193810083 193810324 95007276 46361176 66816517 66808675 66817142 66805559 66800699 66802566 67481173 67469347 67465998 159107882 159109911 159118292 157866296 157872788 157877688 68069041 68065351 68071495 70951061 70923200 70932841 70944487 70953694 86171431 124810112 124512222 124513408 156097697 156096164 156100079 156081857 156101091 83286182 82596506 82541565 83282223 118352498 118380025 118386473 146186117 146182318 71030468 71030930 71027585 71029168 71029466 71755079 72390541 71748176 156083837 156085224 156088773 156088817 156089121 67609547 67587638 67594615 126644823 126649165 66360066 167388342 167383568 167376921 154333954 154341971 154346532 146080950 146094104 146104939 167516418 167520011 167520842 167522938 145479761 145517849 145488952 145492782 84994908 84995376 84997031 84997307 85000765 123473252 123437181 123477120 123975502 71397087 71405862 71656320 71412769

COG0522

110810219 7353 5923991 160688815 1353747 160688766 5231238 10399 4958911 2350964 62176199 5442406 134059461 134064766 134066102 134067367 134072098 134073848 10178677 23497657 23504687 193809685 193810331 2258356 95007412 75756339 187479911 31322456 66476134 50057489 51339699 115443574 161899239 67593230 30468054 11465479 11467355 66815743 7525008 66804533 187764107 71842275 67482301 67481745 183235641 67470430 67484266 67483353 67481871 11467040 12545452 84508556 159112220 159116801 51210012 11467651 162605936 189095326 21449991 157864530 157873760 157876707 11466639 11467476 118411052 68069555 68070681 70940074 124506321 124810096 156100093 156095444 83286205 82540109 60117135 11465685 90994410 15150718 11466526 149071984 11466566 50261300 118356827 146163443 118411185 74325180 71027611 71028602 71025925 74025766 72392295 71747500 156083809 156084614 67595276 66362650 126654511 31442364 167389425 167383739 167395168 167382113 167392132 167388875 160331211 186920132 154332394 154342977 154345646 146077487 146096201 146103077 167526531 167520774 124088436 145505215 145507412 145544819 145546530 145527674 145474345 145487528 145497871 84996521 85000795 11496535 123454657 123397501 123500006 123447382 123381003 123509886 123487896 123468452 71414539 71659082 71420956 71657652

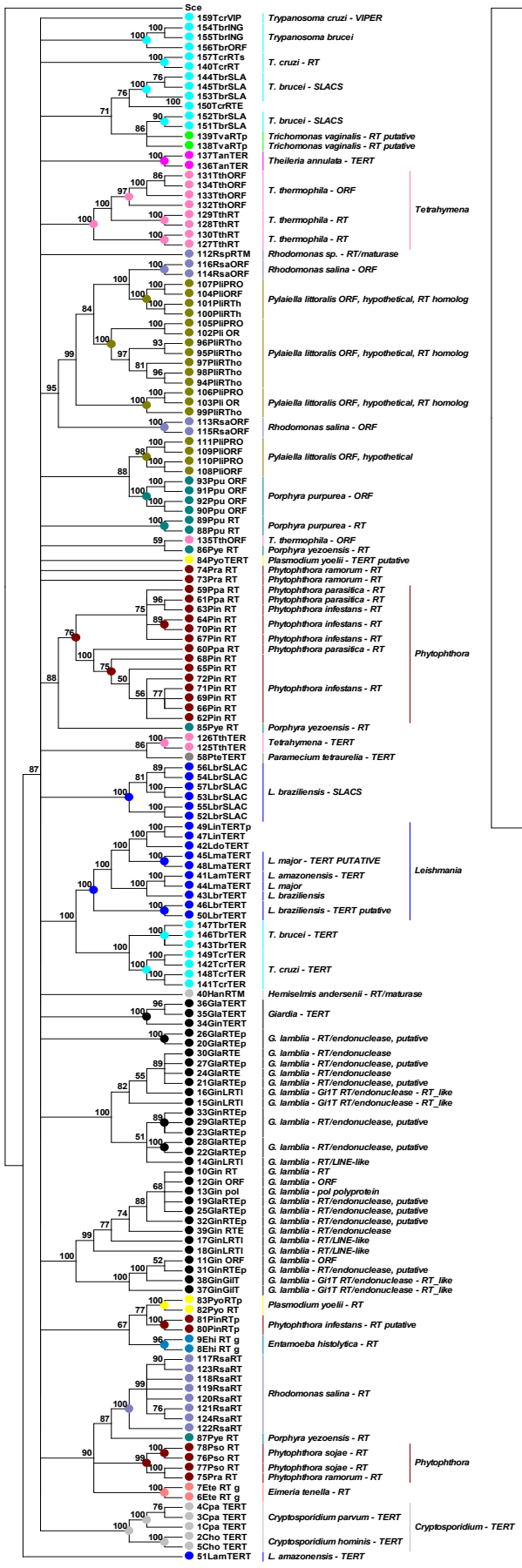
COG0525

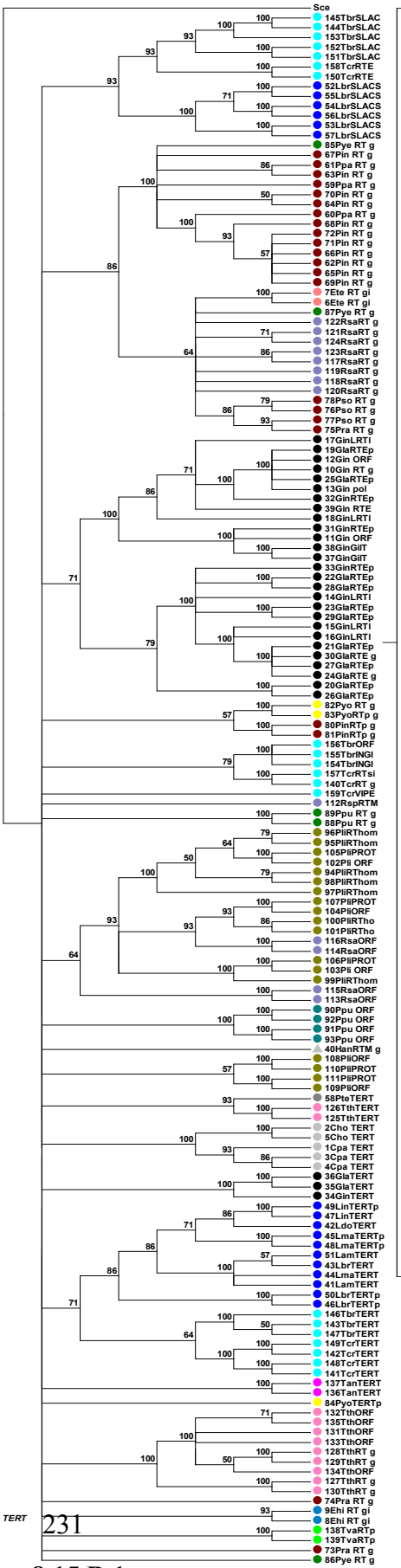
1595807 861093 861094 598414 28828225 57157207 57157209 3219290 57157213 51968325 3219292 3219294 4432760 62358810 57157211 15487676 134064262 134071527 23497663 3764016 193809005 193810324 66817142 66805559 67469347 183235653 159118292 157872788 68071495 70944487 124504885 124810112 156100079 82541565 146186117 71029466 72390541 156089121 67609547 66360066 167383568 154341971 146094104 167527031 167515854 145492782 84997307 123492419 123477120 123478984 71408279 71402226 71656320

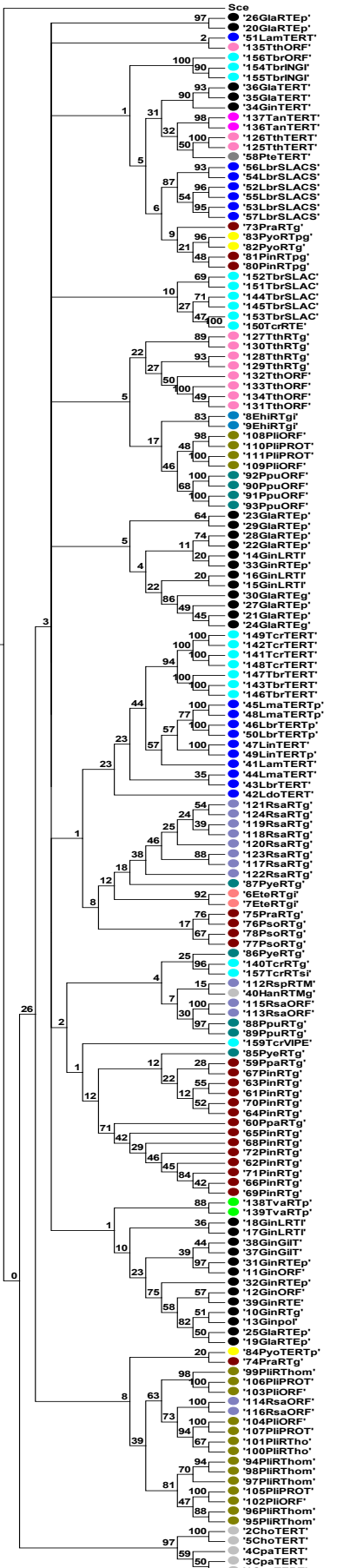
COG0533

62359890 134064340 134071637 23495165 6562727 194247052 193808284 32398931 66827477 66802508 67473009 159115087 157872945 68074493 68062955 68068061 70950864 70935027 124505323 124802749 156094593 156081943 82541770 82540307 146186200 71028570 71029692 72391952 156084680 156089321 67599041 126649333 167379283 154342126 146094427 167517443 167522381 145544082 145536540 84996483 84997521 154422416 71402413 71401774 71422216

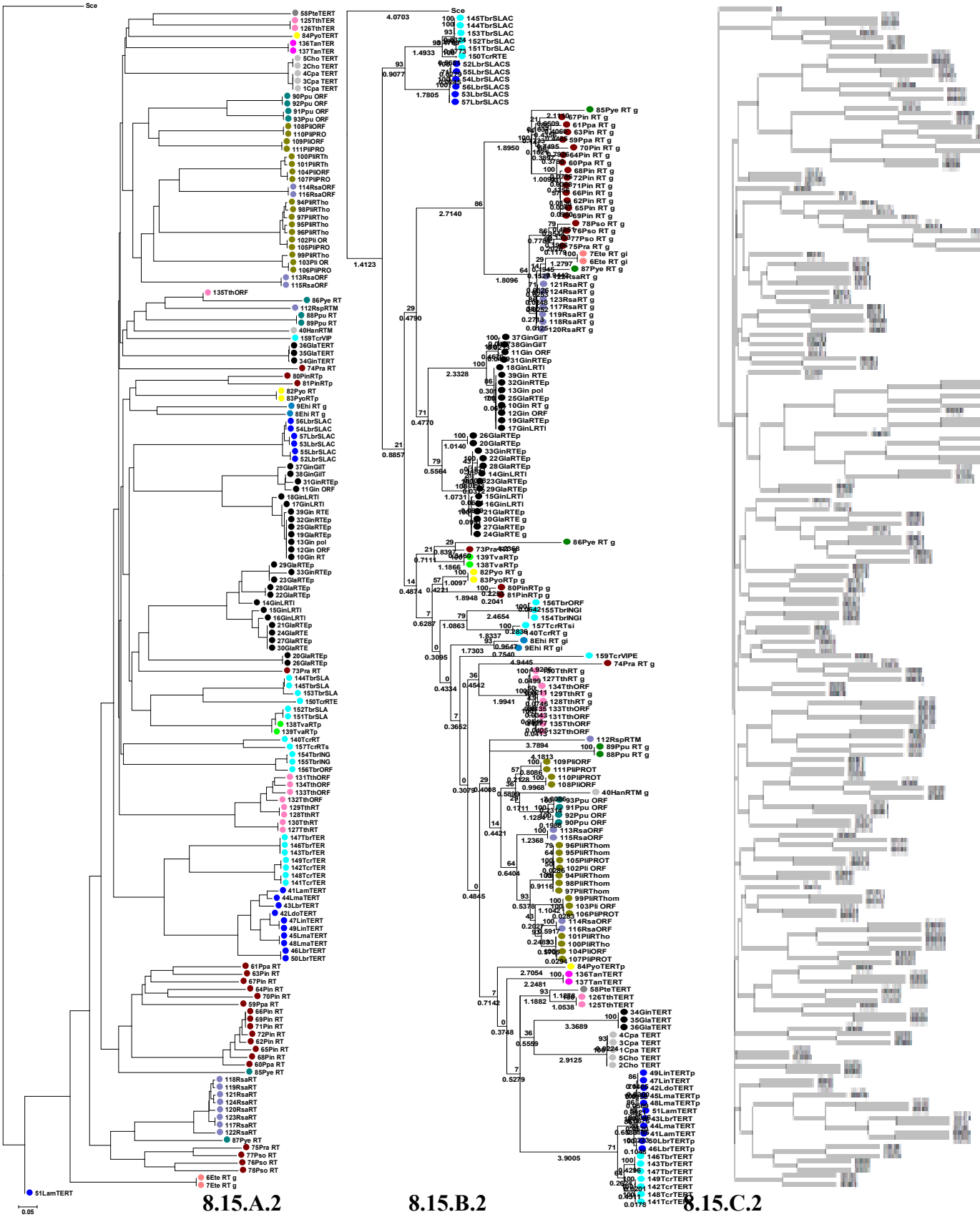
8.4 Filogenômica dos EGM em protozoários







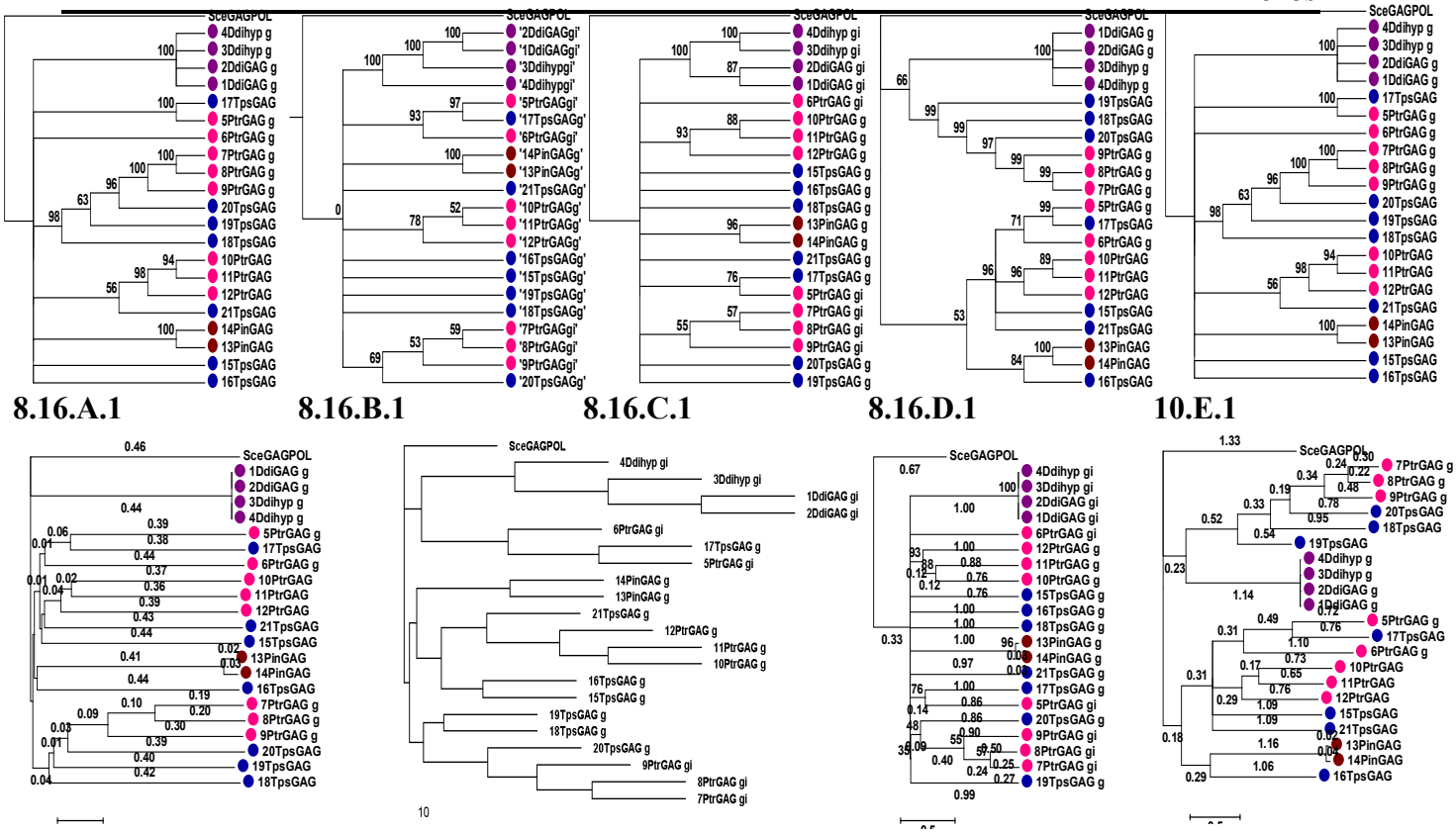
8.15.A.1 8.15.B.1 8.15.C.1



8.15.A.2 8.15.B.2 8.15.C.2

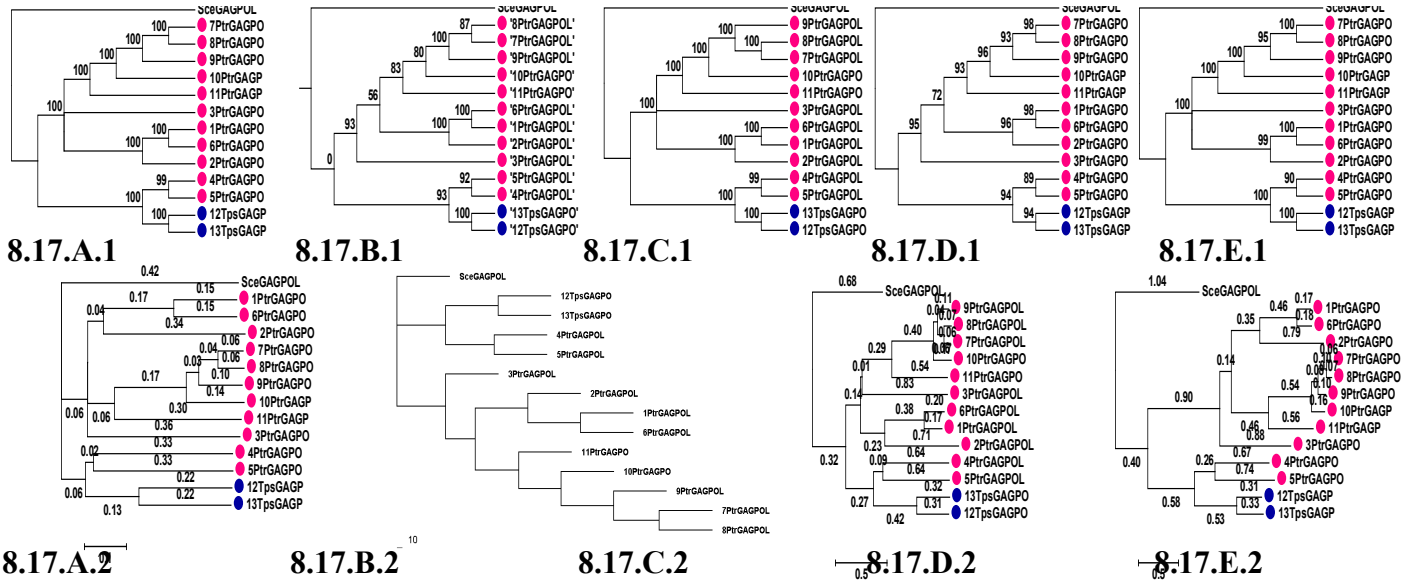
Anexo 8.15 - A filogenia da transcriptase reversa

- A árvore 8.15.A foi construída com o PAUP-AV, a 8.15.B com o RAXML e a 8.15.C com o WEIGHBOR. Todas as árvores foram enraizadas com *S. cerevisiae*.
- As árvores 8.15.A.1, 8.15.B.1, 8.15.C.1 mostram a topologia com valores de *bootstrap*.
- As árvores 8.15.A.2, 8.15.B.2, 8.15.C.2 mostram o comprimento dos ramos.
- As árvores 8.15.B e 8.15.C usaram a matriz do modelo evolutivo WAG e distribuição gama.



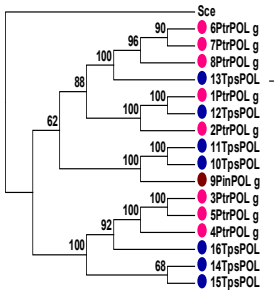
Anexo 8.16 - A filogenia da proteína gag

- A árvore 8.16.A foi construída com o PAUP-AV, a 8.16.B com o WEIGHBOR, a 8.16.C com o PHYML, a 8.16.D com o MRBAYES e a 8.16.E com o PAUP-MP.
- Todas as árvores foram enraizadas com *S. cerevisiae*.
- As árvores 8.16.A.1, 8.16.B.1, 8.16.C.1, 8.16.D.1, 8.16.E.1 mostram valores de *bootstrap*.
- As árvores 8.16.A.2, 8.16.B.2, 8.16.C.2, 8.16.D.2 mostram o comprimento dos ramos.
- As árvores 8.16.B, 8.16.C e 8.16.D usaram a matriz do modelo evolutivo WAG.

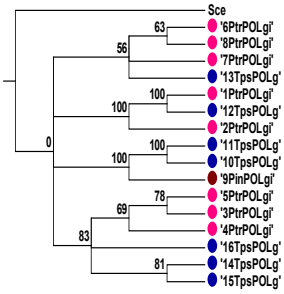


Anexo 8.17 - A filogenia da proteína gag-pol

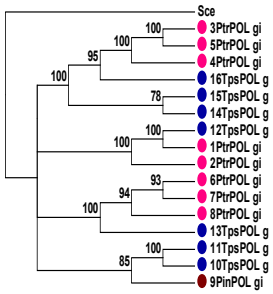
- A árvore 8.17.A foi construída com o PAUP-AV, a 8.17.B com o WEIGHBOR, a 8.17.C com o PHYML, a 8.17.D com o MRBAYES e a 8.17.E com o PAUP-MP.
- Todas as árvores foram enraizadas com *S. cerevisiae*.
- As árvores 8.17.A.1, 8.17.B.1, 8.17.C.1, 8.17.D.1, 8.17.E.1 mostram valores de *bootstrap*.
- As árvores 8.17.A.2, 8.17.B.2, 8.17.C.2, 8.17.D.2 mostram o comprimento dos ramos..
- As árvores 8.17.B, 8.17.C e 8.17.D usaram a matriz do modelo evolutivo WAG.



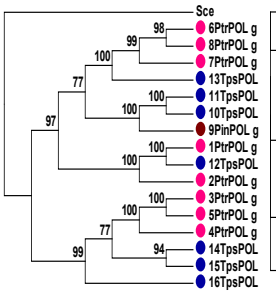
8.18.A.1



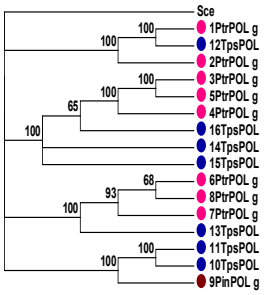
8.18.B.1



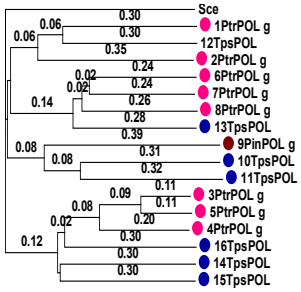
8.18.C.1



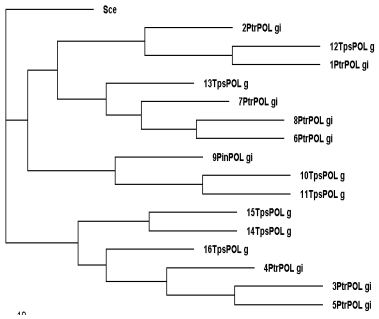
8.18.D.1



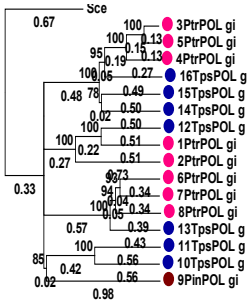
8.18.E.1



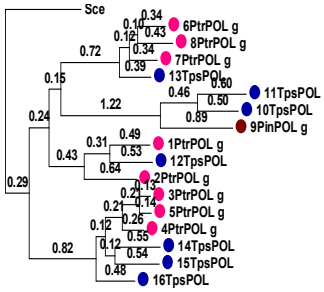
8.18.A.2



8.18.B.2



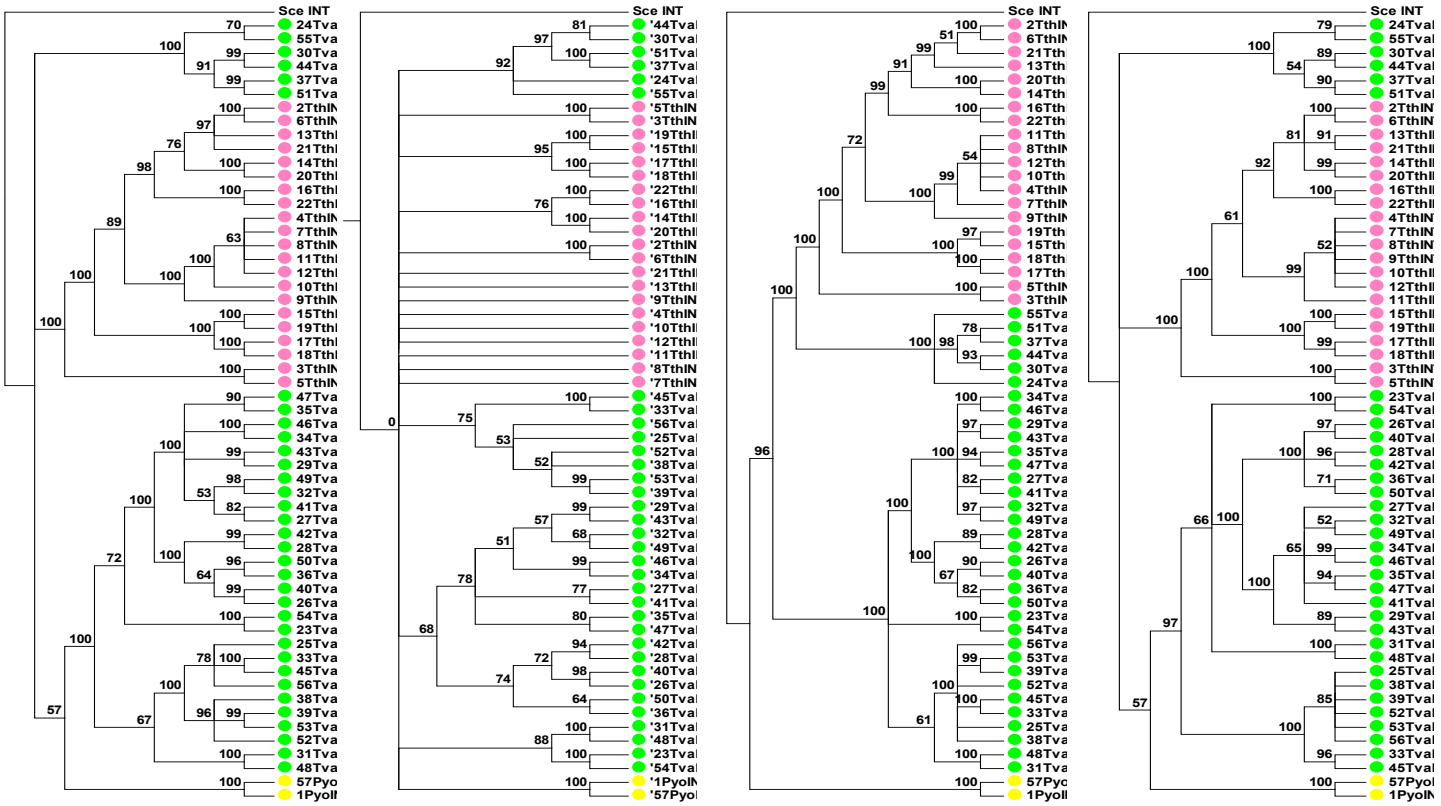
8.18.C.2



8.18.D.2

Anexo 8.18 - A filogenia da proteína pol

- A árvore 8.18.A foi construída com o PAUP-AV, a 8.18.B com o WEIGHBOR, a 8.18.C com o PHYML, a 8.18.D com o MRBAYES e a 8.18.E com o PAUP-MP.
- Todas as árvores foram enraizadas com *S. cerevisiae*.
- As árvores 8.18.A.1, 8.18.B.1, 8.18.C.1, 8.18.D.1, 8.18.E.1 mostram os valores de *bootstrap.*
- As árvores 8.18.A.2, 8.18.B.2, 8.18.C.2, 8.18.D.2 mostram o comprimento dos ramos,.
- As árvores 8.18.B, 8.18.C e 8.18.D usaram a matriz do modelo evolutivo BLOSUM62 e distribuição gama.

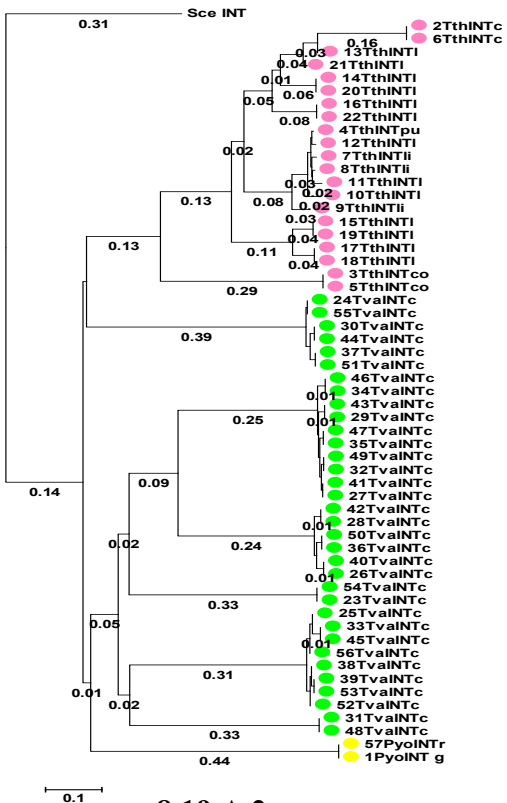


8.19.A.1

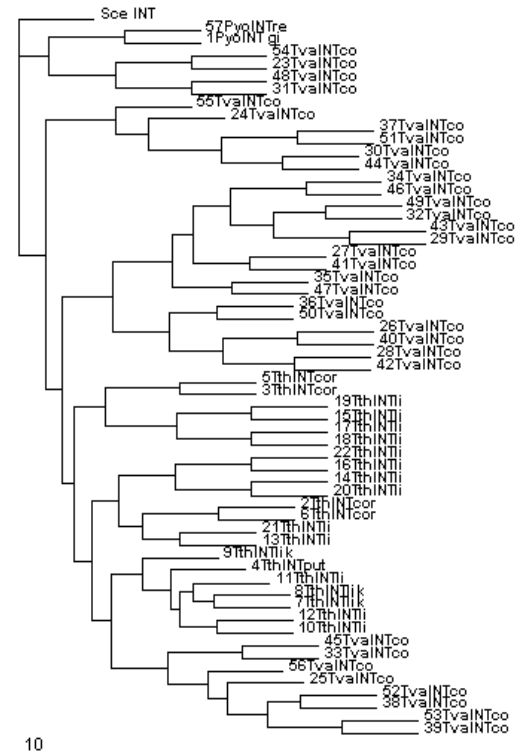
8.19.B.1

8.19.C.1

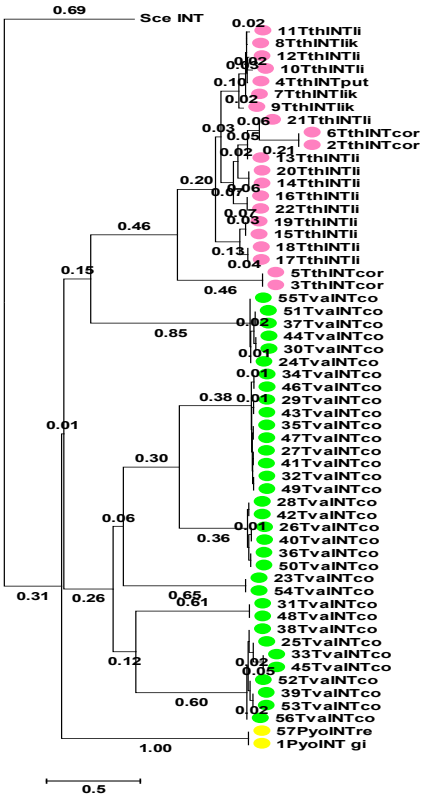
8.19.D.1



8.19.A.2



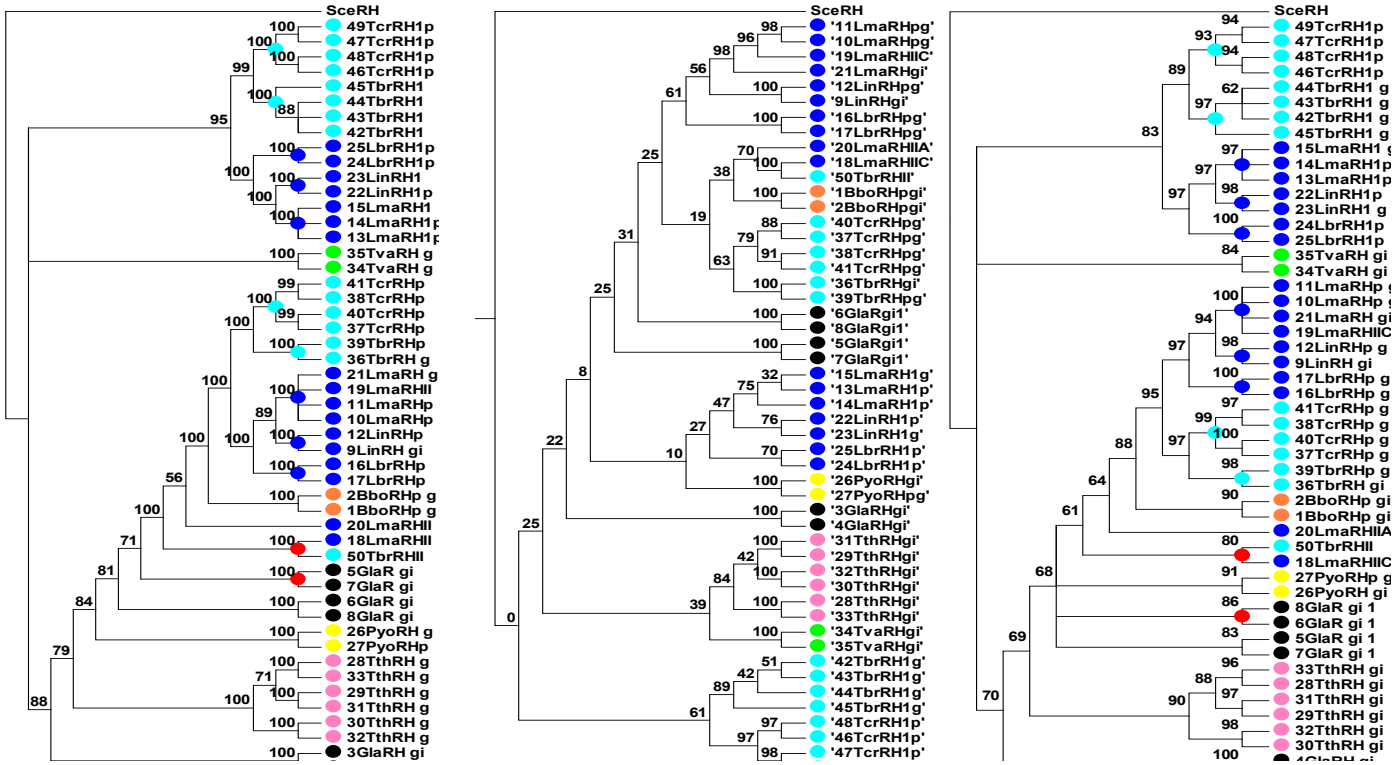
8.19.B.2



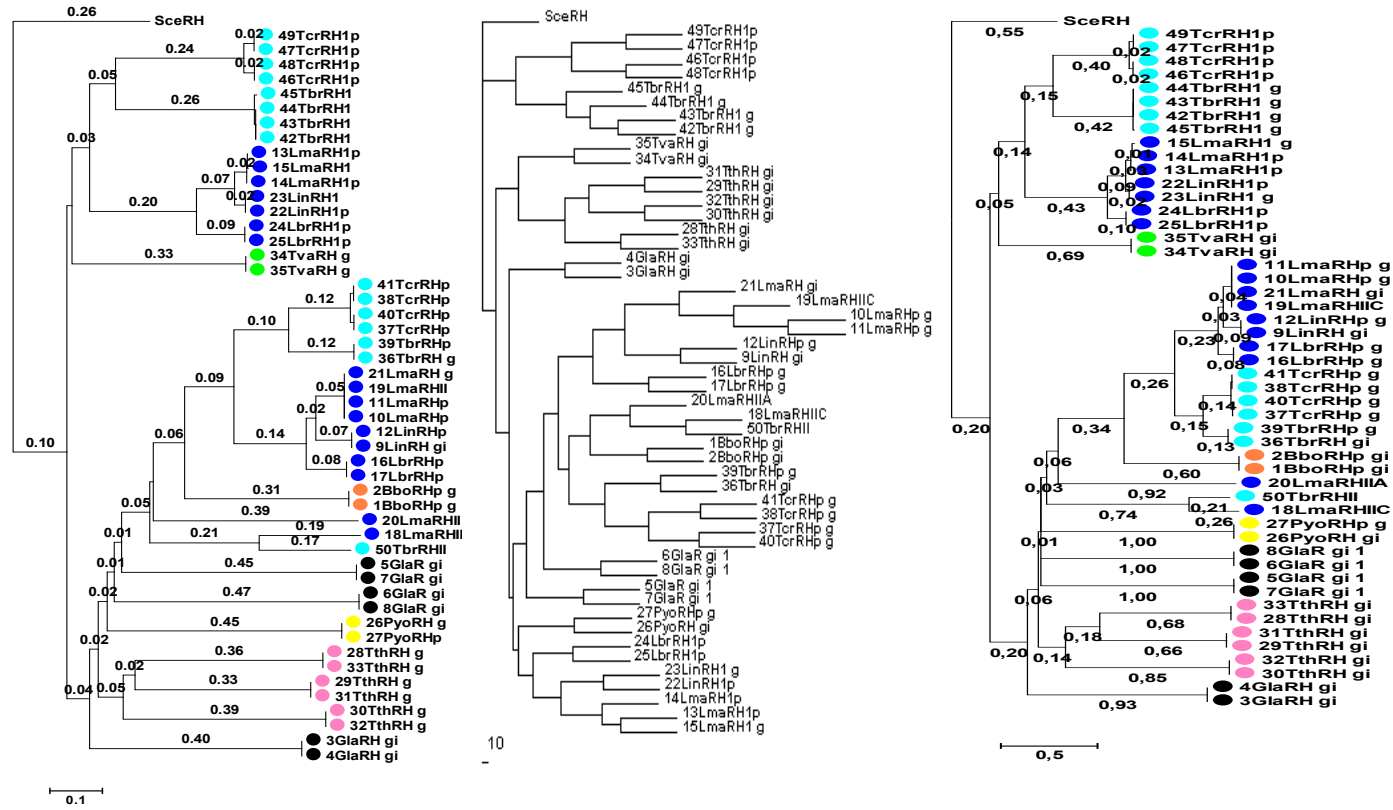
8.19.C.2

Anexo 8.19 - A filogenia da integrase

- A árvore 8.19.A foi construída com o PAUP-AV, a 8.19.B com o WEIGHBOR, a 8.19.C com o PHYML, a 8.19.D com o PAUP-MP.
- Todas as árvores foram enraizadas com *S. cerevisiae*.
- As árvores 8.19.A.1, 8.19.B.1, 8.19.C.1, 8.19.D.1 mostram os valores de *bootstrap.*
- As árvores 8.19.A.2, 8.19.B.2, 8.19.C.2 mostram o comprimento dos ramos.
- As árvores 8.19.B e 8.19.C usaram a matriz do modelo evolutivo BLOSUM62.



8.20.A.1 **8.20.B.1** **8.20.C.1**



8.20.A.2 **8.20.B.2** **8.20.C.2**

Anexo 8.20 - A filogenia da ribonuclease H

- A árvores 8.20.A foram construídas com o PAUP-AV, a 8.20.B com o WEIGHBOR, a 8.20.C com o PHYML, a 8.20.D com PAUP-MP.
- Todas as árvores foram enraizadas com *S. cerevisiae*.
- As árvores 8.20.A.1, 8.20.B.1, 8.20.C.1, 8.20.D.1 mostram os valores de *bootstrap.*
- As árvores 8.20.A.2, 8.20.B.2, 8.20.C.2 mostram o comprimento dos ramos.
- As árvores 8.20.B e 8.20.C usaram a matriz do modelo evolutivo BLOSUM62.