O filoma de *Schistosoma mansoni*

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA GERAL
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

Área de Concentração: Biotecnologia, Genômica e Bioinformática.

DISSERTAÇÃO DE MESTRADO

O FILOMA DE *Schistosoma mansoni*

Orientada: Larissa Lopes Silva
Orientador: Dr. Guilherme Corrêa de Oliveira
Co-orientador: Dr. Juan Antonio Gabaldón Estevan
Co-orientadora: Dra. Laila Alves Nahum

Belo Horizonte
2010

Larissa Lopes Silva

O FILOMA DE *Schistosoma mansoni*

Dissertação apresentada ao Curso de Pós-Graduação em Genética do Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Genética.

Orientador: Dr. Guilherme Corrêa de Oliveira
Co-orientador: Dr. Juan Antonio Gabaldón Estevan
Co-orientadora: Dra. Laila Alves Nahum

Belo Horizonte
2010

043 Silva, Larissa Lopes.
O filoma de schistosoma mansoni. [manuscrito] / Larissa Lopes Silva. – 2010.
71 f. : il. ; 29,5 cm.

Orientador: Guilherme Corrêa de Oliveira. Co-orientadores: Juan Antonio Gabaldón Estevan, Laila Alves Nahum.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Departamento de Biologia Geral.

1. Schistosoma mansoni - Teses. 2. Esquistossomose - Teses. 3. Anotação funcional. 4. Filogenômica. I. Oliveira, Guilherme Corrêa de. II. Estevan, Juan Antonio Gabaldón. III. Nahum, Laila Alves. IV. Universidade Federal de Minas Gerais.Departamento de Biologia Geral. V. Título.

CDU: 575

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

O FILOMA DE *Schistosoma mansoni*

por

Larissa Lopes Silva

Este trabalho foi realizado no Laboratório de Parasitologia Celular e Molecular do Centro de Pesquisas René Rachou[1] e no Centro de Excelência em Bioinformática[2] sob orientação do Dr. Guilherme Corrêa de Oliveira e Co-orientação da Dra. Laila Alves Nahum. Parte das análises computacionais foram executadas em colaboração com o Laboratório de Genômica Comparativa do Centro de Regulação Genômica[3], sob co-orientação do Dr. Juan Antonio Gabaldón Estevan.

1 - Laboratório de Parasitologia Celular e Molecular, Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz – FIOCRUZ, Belo Horizonte, MG, Brasil.

2 - Centro de Excelência em Bioinformática, FIOCRUZ, Belo Horizonte, MG, Brasil.

3 - Comparative Genomics, Center for Genomic Regulation, Barcelona, Espanha.

Belo Horizonte

2010

Não importa o que aconteça, o que importa é como você reage. Eu sou um Bambu. O vento me balança, me dobra, mas não me quebro de jeito nenhum.

Laila Nahum

Dedico este trabalho à minha família que me ensinou a sonhar, a buscar, que me fez forte para alcançar, que me ensinou a aceitar os outros como são e a exigir de mim o melhor que posso ser.

AGRADECIMENTOS

Agradeço ao meu orientador Dr. Guilherme Oliveira, por inúmeras oportunidades a mim concedidas, pela confiança, paciência, conhecimento compartilhado e pelo convívio com esta pessoa justa, gentil, sábia, bem humorada, extremamente educada e tão otimista que transmite paz de espírito e torna impossível imaginar que algo não vai dar certo com ele por perto.

À minha co-orientadora Dra. Laila Nahum, fonte de sabedoria, uma das gratas surpresas que a Casa Amarela reservou para mim. Meses depois de conhecê-la e me tornar profunda admiradora do seu trabalho, tive a felicidade de ser co-orientada por ela. Além do enorme conhecimento científico, durante longos e prazerosos diálogos me ensinou lições de vida, de determinação e superação.

Ao meu co-orientador Dr. Toni Gabaldón por abraçar este projeto, pelo apoio, incentivo e infra-estrutura compartilhada, sem os quais este trabalho não teria sido realizado.

Ao Programa de Pós-Graduação em Genética pela oportunidade e apoio que tanto contribuíram para minha formação.

Ao Omar e à Roberta do Laboratório de Helmintologia e Malacologia Médica (LHMM) do Centro de Pesquisas René Rachou (CPqRR) que me receberam como aluna de iniciação científica e foram os responsáveis pelo meu ingresso na ciência.

Aos amigos do CEBIO, Anderson, Éric, Fabiano, Francislon, Mari, Núbia, Leandro, Rômulo e Roney pelas conversas agradáveis, momentos divertidos e ajuda incalculável. Em especial, à Ângela e ao Adhemar, sinônimos de competência e disponibilidade em ajudar.

Aos amigos do LHMM, Cristiano Massara, Fernanda, Sandrinha, Solange, Martin e Ricardo, pela convivência, companhia e ajuda constante. Em especial, à Cris Lafetá que tanto contribuiu e contribui para o meu crescimento profissional.

Aos amigos e colegas de trabalho do Laboratório de Parasitologia Celular e Molecular (LPCM) do CPqRR, Luiza, Lívia, Fernanda Ludolf, Flávio, Betânia, Marcilene, Karla, Kelly,

Elis, Regina, Rosiane, Policarpo, Rosana, Jerônimo, Raul, Patrícia, Silvane e Fernanda Souto por tornarem esta jornada super agradável.

Às famílias Scholte e Carvalho por terem me acolhido com tanto carinho. Em especial, à linda e charmosa Vó Ruth, pelo apoio, incentivo e torcida.

Às famílias Lopes e Silva, por tudo.

Aos amigos Kelly e Márcio. Adoro vocês!

Ao Rô pelo amor, paciência, companheirismo, conselhos e por tornar minha vida ainda mais feliz.

Às minhas mães Cléa e Zélia, ao meu pai Márcio, às minhas irmãs Érika, Natália e Giselle, ao meu irmão Marcinho e ao Etê, pelo incentivo constante, por me fazerem acreditar na vida, nas pessoas, no amor e por serem parte de mim e permitirem que eu seja parte de vocês.

A todos vocês meu sincero e carinhoso MUITO OBRIGADA!

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE QUADROS

LISTA DE ABREVIATURAS

AA: aminoácidos
AIC: Akaike Information Criterium, Critério de Informação de Akaike
ATP: Adenosine Triphosphatase; Adenosina Trisfosfatase; Trifosfatase de Adenosina
aLRT: approximate likelihood ratio test;
COX: cytochrome *c* oxidase; citocromo oxidase *c*
CYTB: Cytochrome *b;* citocromo *b*
DB: Database; Banco de Dados
DNA: Deoxyribonucleic acid; ácido desoxirribonucléico
ETE: Enviroment for Tree Exploration; Ambiente para Exploração de Árvores
GO: Gene Ontology
JGI: Joint Genome Institute
Mb: megabase
ML: Maximum Likelihood; Máxima Verossimilhança
MUSCLE: Multiple Sequence Comparison by Log-Expectation
NCBI: National Center for Biotechnology Information
ND: Nicotinamide Dehydrogenase; Nicotinamida Desidrogenase
NJ: Neighbor Joining; Agrupamento de Vizinhos
pb: pares de base
PhylomeDB: Phylome Database
RNA: Ribonucleic acid; ácido ribonucléico
SchistoDB: Schistosoma mansoni Database
SilkDB: Silkworm Genome Database
SGBD: sistema gestor de base de dados
SQL: structured query language

RESUMO

Schistosoma mansoni é um parasito platelminto da classe Trematoda e é um dos agentes etiológicos da esquistossomose, uma das doenças infecciosas e parasitárias mais prevalentes no mundo, sendo endêmica em 76 países e territórios onde aproximadamente 210 milhões de pessoas encontram-se infectadas e outras 779 mil vivem em áreas de risco de infecção. O proteoma predito deste parasito foi publicado recentemente e contém mais de 13.000 proteínas. Contudo, mais de 40% permanece sem função predita ou caracterização experimental. Em virtude do exposto, o presente trabalho teve como objetivo principal aprimorar a anotação do genoma de *S. mansoni* utilizando a abordagem filogenética para a predição de ortólogos. A reconstrução da história evolutiva de todas as proteínas codificadas no genoma de *S. mansoni* foi realizada através de um *pipeline* automático, como implementado no *PhylomeDB* (http://phylomedb.org), um banco de dados de filomas. O filoma resultante possui 8.818 árvores filogenéticas obtidas a partir da análise comparativa de 13.285 proteínas de *S. mansoni* com os potenciais homólogos em 16 outros organismos, incluindo uma planta, fungos, nematóides, artrópodes, urocordados, cefalocordados e vertebrados. Utilizando a abordagem filogenética para a predição de ortólogos foi possível transferir anotação funcional (termos GO) para 5.507 proteínas de *S. mansoni*, das quais 946 eram anotadas previamente como “proteínas hipotéticas” ou “proteínas expressas” que correspondem a proteínas cuja função é completamente desconhecida. Além de promover uma melhora significativa na predição funcional do proteoma de *S. mansoni*, os resultados desta análise provêm informações importantes sobre a evolução do genoma deste parasito, como a identificação de duplicações gênicas que podem estar relacionadas a especificidades morfológicas ou fisiológicas deste organismo. Todos os alinhamentos de sequências, árvores filogenéticas e anotação funcional serão disponibilizadas através dos bancos de dados *PhylomeDB* (http://phylomedb.org) e *SchistoDB* (http://schistodb.net), fornecendo uma vasta fonte de informações para a comunidade científica.

ABSTRACT

Schistosoma mansoni is a Platyhelminth parasite from the Trematode class. It is one of the etiologic agents of human schistosomiasis, a tropical neglected disease that is endemic in 76 countries where 210 million people are infected and other 779.000 live in areas of risk of infection. The parasite predicted proteome was recently published and contains over 13.000 proteins. However, more than 40% remain without a predicted function or experimental characterization. Therefore, the present study had as main goal to improve the *S. mansoni* functional annotation using the phylogenetic approach for the prediction of orthologues. The reconstruction of the evolutionary histories of all proteins encoded in the *S. mansoni* genome was performed through an automatic pipeline, as implemented in PhylomeDB (http://phylomedb.org), a phylome database. The resulting phylome has 8.818 phylogenetic trees derived from the comparative analyses of 13.285 *S. mansoni* proteins and their potential homologs in 16 other organisms, including a plant, fungi, nematodes, arthropods, urochordates, cephalochordates and vertebrates. Using this phylogeny-based approach, we could transfer functional annotations from Gene Ontology (GO) to 5.507 *S. mansoni* proteins, 946 of wich were previously annotated as “hypothetical” or “expressed protein”, corresponding to proteins with a completely unknown function. Besides promoting a significant improvement in the functional prediction of the *S. mansoni* proteome, this approach has provided useful information about the parasite’s genome evolution, such as identification of gene duplications that may be related to morphological or physiological characteristics of this organism. All sequence alignments, phylogenetic trees, and functional annotation will be made publicly available through PhylomeDB (http://phylomedb.org) and *SchistoDB* (http://schistodb.net) databases providing a powerful resource for the scientific community.

1. INTRODUÇÃO

1.1. A Esquistossomose

1.1.1. Aspectos gerais e epidemiológicos da doença

A esquistossomose é uma das doenças infecciosas e parasitárias mais prevalentes no mundo, sendo endêmica em 76 países e territórios onde aproximadamente 210 milhões de pessoas encontram-se infectadas e outras 779 mil vivem em áreas de risco de infecção (Engels *et al.*, 2002; Steinmann *et al*., 2006). De todos os países com relatos de transmissão, 46 estão na África onde residem cerca de 97% do total de infectados e 85% da população mundial em risco de infecção (Steinmann *et al*., 2006). Apenas na África sub-saariana, aproximadamente 280.000 mortes por ano são atribuídas à esquistossomose (van der Werf *et al.*, 2003).

Esta doença tem como agentes etiológicos platelmintos trematodas do gênero *Schistosoma*, sendo *S. japonicum*, *S. haematobium* e *S. mansoni* as principais espécies que acometem humanos (Han *et al*., 2009). O *S. mansoni* ocorre em grande parte da África sub-saariana, regiões do Oriente Médio, Venezuela, Caribe e é a espécie responsável pela ocorrência da esquistossomose no Brasil (Gryseels *et al*., 2006) (Figura 1).

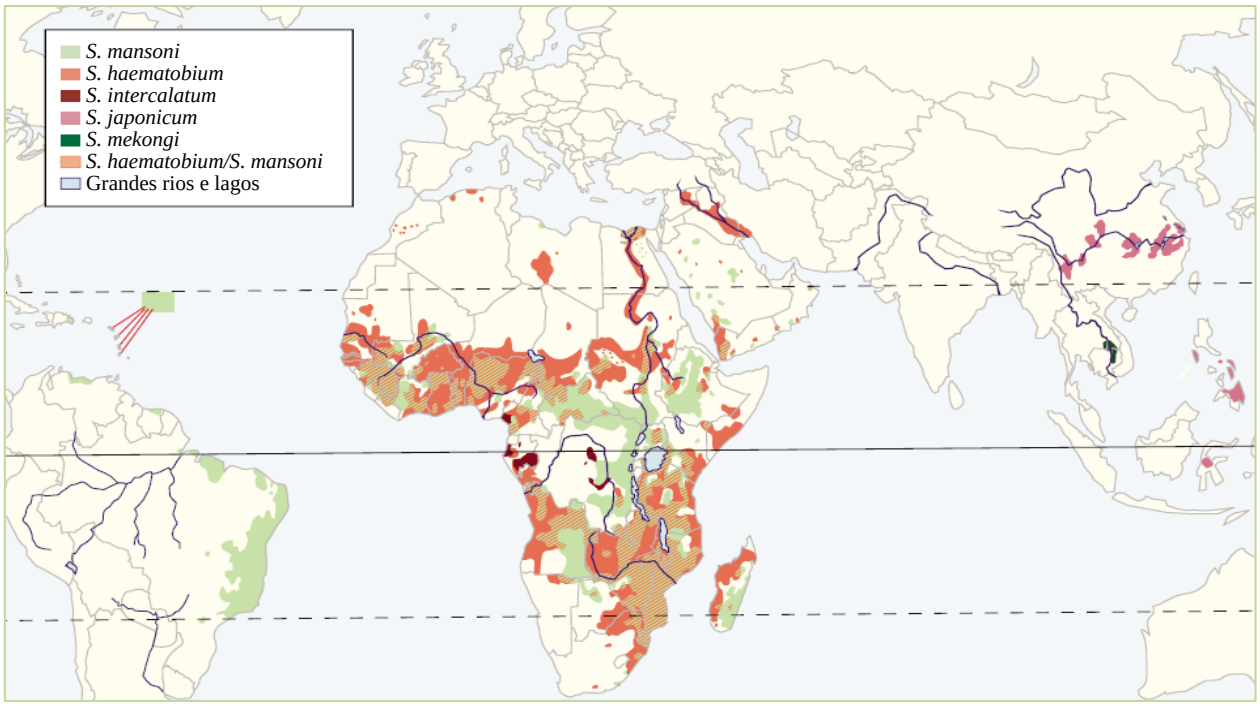


Figura 1: Distribuição global da esquistossomose. *S. mansoni* possui ampla distribuição na África Subsaariana, nordeste do Brasil, Suriname, Venezuela, Caribe e possui distribuição restrita no Egito e Arábia (Adaptado de Gryseels *et al*., 2006).

Em função da esquistossomose ser uma doença veiculada pela água, esta acomete principalmente populações pobres que vivem próximo a coleções hídricas contaminadas pelo parasito e que não têm acesso a cuidados adequados ou medidas de prevenção eficazes (Alam *et al.*, 2009). Na maioria das vezes, a água oriunda destas coleções é utilizada em atividades diárias como lavar, cozinhar e para fins de geração de renda como agricultura e pesca. Taxas de infecção e prevalência mais altas são encontradas em crianças em idade escolar, adolescentes e adultos jovens, afetando negativamente o desempenho escolar e comprometendo o desenvolvimento social e econômico em áreas fortemente afetadas (Steinmann *et al*., 2006; Hotez & Kamath, 2009).

Atualmente o controle da esquistossomose depende principalmente do tratamento de pacientes infectados, utilizando praziquantel, a única droga disponível para tratamento (TDR, 2007; Bruun & Aagaard-Hansen, 2008; Berriman *et al.*, 2009). Entretanto, esta droga não previne contra a reinfecção e a eficácia é variável dependendo do sexo do parasito, estágios de desenvolvimento e período de infecção, além de evidências de parasitos resistentes a este medicamento terem sido obtidas em laboratório e no campo (Liang *et al.*, 2003; Pica-Mattoccia & Cioli, 2004; Botros & Bennet, 2007; Melman *et al*., 2009).

Embora a distribuição da esquistossomose tenha mudado nos últimos 50 anos com o sucesso de alguns programas de controle, o número estimado de pessoas infectadas ou em risco de infecção não foi alterado, uma vez que resultados positivos das medidas de controle foram obtidos em áreas onde o número de pessoas infectadas e em risco de infecção é muito pequeno (Engels *et al.*, 2003).

Em virtude do exposto, a esquistossomose continua sendo uma das doenças infecciosas e parasitárias mais prevalentes no mundo, responsável por perdas econômicas significativas e sérias consequências para a saúde pública.

1.1.2. O parasito *Schistosoma mansoni*

Schistosoma mansoni (Figura 2) é um parasito lofotrocozoário, platelminto da classe Trematoda e é um dos agentes etiológicos da esquistossomose intestinal. O nome da categoria taxonômica Lophotrochozoa deriva de características morfológicas dos membros do grupo – “Lopho” faz referência à estrutura alimentar ciliada presente, por exemplo, em representantes dos filos Brachiopoda e Phoronida e “trochozoa” se refere à larva trocófora, estágio larval presente em representantes dos filos Mollusca, Annelida, Platyhelminthes, dentre outros (Telford, 2006). Os vermes *S. mansoni* adultos possuem coloração branca ou acinzentada, os machos medem aproximadamente 1 cm e as fêmeas 1,5 cm, têm um tegumento complexo e duas ventosas, sendo uma oral e outra ventral (Melo & Coelho, 2005;

Gryseels *et al*., 2006). Ao contrário de outros trematódeos, estes organismos são dióicos. O macho possui um canal ginecóforo constituído por dobras no sentido longitudinal onde ele aloja e fecunda a fêmea (Melo & Coelho, 2005; Gryseels *et al*., 2006).

Figura 2: Microscopia eletrônica de um casal de parasitos coloridos artificialmente. A fêmea (em verde) está localizada no interior do canal ginecóforo do macho (em marrom). Disponível em http://emtrix.dbs.umt.edu/imagery/imagery.htm. Acesso em: 30 mar. 2010.

O ciclo biológico de *S. mansoni* apresenta uma alternância de gerações entre os hospedeiros intermediários, moluscos do gênero *Biomphalaria* (Gryseels,*et al.*, 2006), e os hospedeiros definitivos vertebrados, roedores, primatas e o homem (Figura 3) (Han *et al.*, 2009). Ao longo do ciclo, complexas adaptações morfológicas e fisiológicas são necessárias para que este parasito consiga sobreviver em ambientes completamente distintos como a água e o meio interno de seus hospedeiros (LoVerde *et al.*, 2004; Melo & Coelho, 2005; Webster *et al.*, 2010).

A transmissão do *S. mansoni* ocorre através do contato com a água onde cercárias infectantes nadam em busca de um hospedeiro vertebrado (Ross *et al.*, 2002). Embora as cercárias possam viver por 36 a 48 horas, sua maior atividade e capacidade infectiva ocorrem nas primeiras oito horas de vida (Melo & Coelho, 2005). Uma vez em contato com a pele do homem, as cercárias se fixam e, através de movimentos vibratórios intensos,

promovem a penetração do corpo cercariano com concomitante perda da cauda (Melo & Coelho, 2005). Este processo tem duração aproximada de 5 a 15 minutos.

Após a penetração, as larvas resultantes, denominadas esquistossômulos, adaptam-se às condições fisiológicas do meio interno, migram pelo tecido subcutâneo e, ao penetrarem em um vaso sanguíneo, são levadas passivamente da pele até os pulmões pelo sistema vascular, via lado direito do coração (Ross *et al.*, 2002). Dos pulmões, os esquistossômulos se dirigem para o sistema porta-hepático. Uma vez no interior deste sistema, os esquistossômulos se alimentam e se desenvolvem transformando-se em machos e fêmeas 25-28 dias após a penetração (Ross *et al.*, 2002; Melo & Coelho, 2005). Ainda no sistema porta eles se acasalam e os casais migram para a veia mesentérica inferior onde farão oviposição.

Os ovos imaturos são depositados nos tecidos em torno do 35º dia da infecção e a formação do miracídio (ovo maduro) demanda, aproximadamente, seis dias. Cada fêmea põe cerca de 400 ovos por dia que se depositam na parede dos capilares e vênulas. Cerca de 50% destes são expulsos ativamente para a luz do intestino e são eliminados juntos com as fezes. Os ovos que não são eliminados podem ficar retidos na mucosa intestinal ou serem arrastados pela corrente circulatória podendo se alojar em vários órgãos, principalmente no fígado. Os primeiros ovos são vistos nas fezes cerca de 42 dias após a infecção do hospedeiro (Melo & Coelho, 2005). Os ovos eliminados com o bolo fecal têm uma expectativa de vida de 24 horas (fezes líquidas) a cinco dias (fezes sólidas). Alcançando a água, os ovos liberam o miracídio, estimulado pela temperatura, luz intensa e oxigenação da água (Melo & Coelho, 2005). O miracídio então nada em busca de hospedeiros intermediários, moluscos do gênero *Biomphalaria*. A capacidade de penetração restringe-se a cerca de oito horas após a eclosão. Em contato com o molusco, intensos movimentos combinados com a ação enzimática permitem a penetração do miracídio nos tecidos do hospedeiro invertebrado. No interior do molusco, o miracídio transforma-se em esporocisto, sofre reprodução assexuada e os esporocistos transformam-se em cercárias. Cada miracídio pode gerar de 300 a 400 mil cercárias (Ross *et al.*, 2002; Melo & Coelho, 2005).

A esquistossomose mansônica é basicamente uma doença que decorre da resposta inflamatória granulomatosa que ocorre em torno dos ovos vivos do parasito. Entre 50 e 120 dias após a infecção caracteriza-se a fase aguda da doença onde é comum observar sintomas como febre, sudorese, calafrios, emagrecimento, fenômenos alérgicos, diarréia, disenteria, cólicas e etc. O paciente não tratado poderá evoluir para a fase crônica da doença com o desenvolvimento de hepatomegalia, esplenomegalia, varizes e ascite (barriga d'água). Estudos indicam que o período de vida médio de *S. mansoni* é de cinco anos, entretanto, relatos comprovam a existência de casais eliminando ovos por mais de trinta anos no interior do hospedeiro vertebrado (Melo & Coelho, 2005).

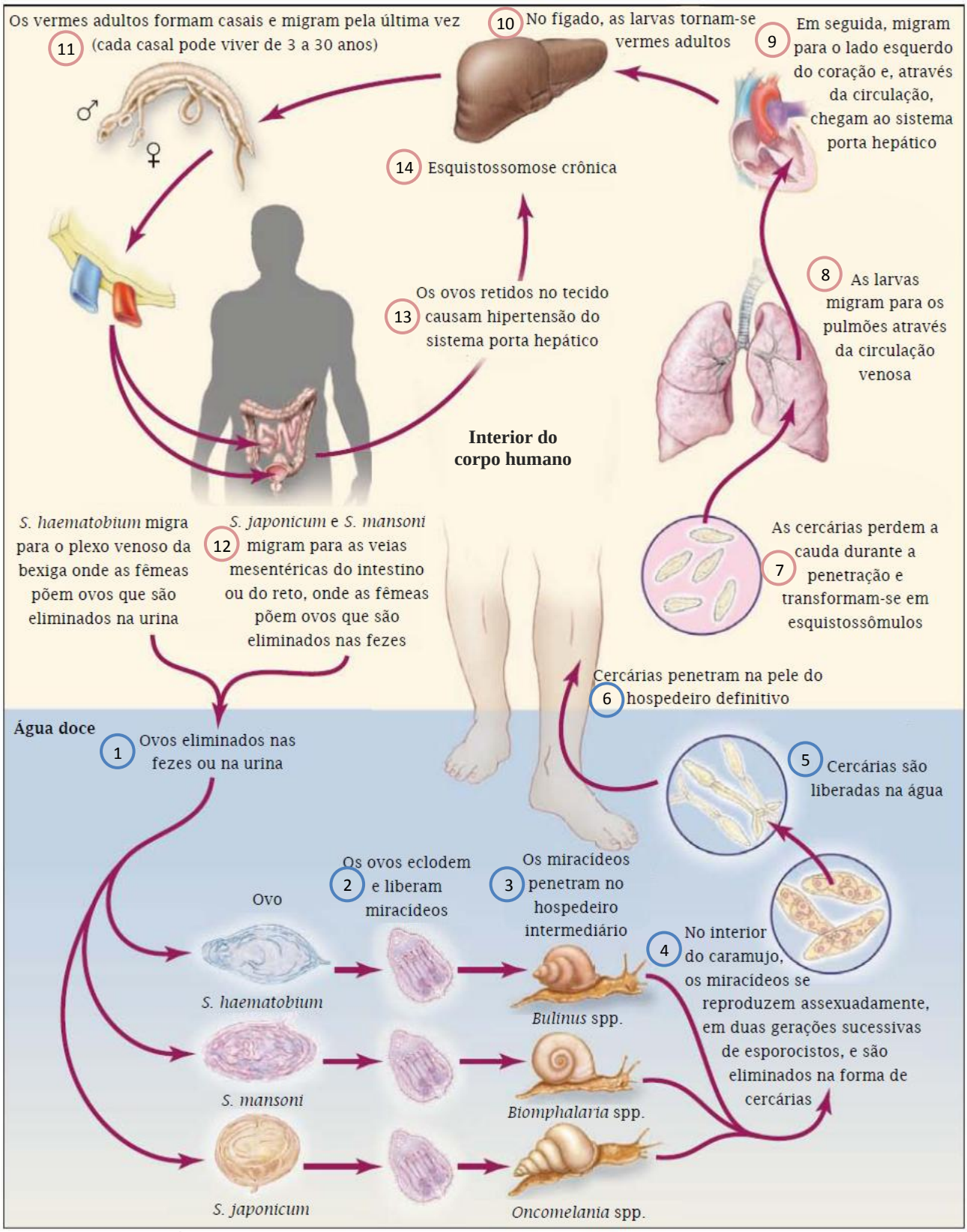


Figura 3: Ciclo biológico do *Schistosoma mansoni* (Adaptado de Ross *et al.*, 2002). O hospedeiro definitivo é infectado pela cercária quando em contato com coleções hídricas contaminadas (6). No interior do hospedeiro, a cercária transforma-se em esquistossômulo (7) que mais tarde, no sistema porta-hepático, transforma-se em verme adulto (11). Os vermes adultos se acasalam e casais de *S. mansoni* migram para as veias mesentéricas onde a fêmea inicia a oviposição (12). Parte dos ovos atinge a luz do intestino e são eliminados com as fezes (1). Em contato com a água, os ovos eclodem liberando miracídios (2) que nadam em busca de hospedeiros invertebrados (3), moluscos do gênero *Biomphalaria*, no caso de parasitos da espécie *S. mansoni*. No interior do molusco, os miracídios transformam-se em esporocistos que se reproduzem assexuadamente e transformam-se em cercárias infectantes (4).

1.1.3. O genoma e proteoma de *S. mansoni*

Schistosoma mansoni possui um genoma diplóide com 7 pares de cromossomos autossômicos e um par de cromossomos sexuais (Criscione *et al.*, 2009). A fêmea é heterogamética, possuindo o par de cromossomos sexuais ZW e o macho o par ZZ (Short *et al*., 1979; Criscione *et al.*, 2009). O genoma deste parasito foi recentemente publicado compreendendo um total de 363 megabases (Mb) distribuídas em 5.745 *scaffolds* (*supercontigs*) maiores que 2 kilobases (Kb) cada (Berriman *et al.*, 2009). Aproximadamente 70% do tamanho estimado do genoma foi ordenado em cromossomos constituindo o mapa genético de *S. mansoni* (Criscione *et al.,* 2009).

O genoma mitocondrial deste parasito é bastante similar ao dos demais metazoários, possui 14.415 pb e codifica 36 genes: 2 RNA's ribossomais, 22 RNA's transportadores e 12 proteínas (COXI, COXII, COXIII, NAD1, NAD3, NAD4, NAD5, ATP6 e CYTB) (Le *et al.*, 2001). Assim como na maioria dos nematóides e platelmintos, o gene ATP8 é ausente (Hu & Gasser, 2006; Vallès & Boore, 2006).

A versão do genoma utilizada no presente trabalho foi obtida no *SchistoDB 2.0* (http://schistodb.net) e possui um total de 12.003 genes (incluindo os genes mitocondriais) que codificam 13.285 transcritos (Zerlotini *et al.*, 2009), embora o número real de genes possa chegar a 17.500 (Berriman *et al.*, 2009). Entretanto, do proteoma predito até o momento, mais de 42% ainda não possui função conhecida, com proteínas anotadas apenas como "proteína hipotética" (do inglês, *hypothetical protein*) ou "proteína expressa" (do inglês, *expressed protein)*, para as quais termos *Gene Ontology* (GO) não foram identificados. Embora a anotação "*expressed protein"* sugira confirmação de expressão gênica, nos dados referentes ao genoma de *S. mansoni* esta informação não está associada a esta nomenclatura. Assim como "*hypothetical protein"*, "*expressed protein"* quer dizer apenas função desconhecida (do inglês, *unknown*).

A utilização dos termos GO na anotação funcional fornece uma base de dados estruturada para a descrição dinâmica e organizada da função gênica e dos produtos gênicos de qualquer organismo (Ashburner *et al.*, 2000). O projeto GO resulta de um esforço colaborativo para atender à necessidade de descrições consistentes de produtos gênicos em bancos de dados diferentes. O vocabulário padronizado de termos pode ser aplicado a três ontologias independentes: processos biológicos, função molecular e componente celular. Esta metodologia tem sido adotada pela comunidade científica para a anotação funcional de genomas recém sequenciados (Lewis, 2005).

A disponibilidade de informações genômicas permite a aceleração do conhecimento da Biologia de Sistemas de *S. mansoni* e das relações estabelecidas com os hospedeiros,

em especial com o hospedeiro definitivo. Entretanto, um dos desafios que persistem é compreender quais são os genes ganhos, perdidos ou duplicados em *S. mansoni*, em relação ao genoma dos demais organismos, e quais são os padrões de expressão destes genes os quais permitem que este parasito sobreviva por anos em um ambiente potencialmente hostil da circulação sanguínea humana, protegido da ação do sistema imunológico e/ou intervindo ativamente, tornando a resposta do hospedeiro ineficaz.

1.2. Análises filogenéticas

Comparar a enorme variedade de organismos vivos e suas características tem sido uma tarefa central na Biologia ao longo dos séculos, uma vez que acredita-se que todos os organismos vivos estejam relacionados por uma ancestralidade comum (e.g. Descorps-Declère *et al.*, 2008; Page & Holmes, 1998). Com o advento das técnicas de sequenciamento de DNA associadas ao desenvolvimento de bancos de dados e ferramentas computacionais, a reconstrução da história evolutiva baseada em sequências de nucleotídeos e/ou aminoácidos tornou-se possível e necessária (Gabaldón, 2005; Snel *et al.*, 2005; Whelan *et al.*, 2001).

A reconstrução filogenética utilizando sequências moleculares de genes e/ou proteínas, assim como a reconstrução baseada em dados morfolólogicos, envolve a identificação de caracteres homólogos compartilhados pelos organismos, seguida da inferência de árvores filogenéticas a partir da comparação destes caracteres (Figura 4) (Delsuc *et al.*, 2005). Homologia é uma relação existente entre duas entidades biológicas, como por exemplo, duas sequências moleculares ou dois caracteres anatômicos que derivam do mesmo ancestral (Fitch, 1970; Koonin, 2005; Descorps-Declère *et al.*, 2008; Gabaldón, 2008b). De forma mais detalhada, as relações evolutivas são melhores estabelecidas por meio da detecção de ortologia, ao invés de apenas homologia (Huerta-Cepas *et al*., 2007). Ortólogos são proteínas ou genes homólogos derivados de um ancestral comum por especiação, enquanto parálogos são homólogos derivados por duplicação gênica (Fitch, 1970). A detecção de relações de ortologia é idealmente realizada através da identificação de eventos de especiação e duplicação evidenciados pela análise filogenética (Figura 6).

As filogenias são hipóteses evolutivas que indicam relações de ancestralidade representadas através de árvores filogenéticas. Uma árvore é um grafo utilizado para representar a história evolutiva de um grupo de moléculas ou organismos e consiste de nós conectados por ramos (Page & Holmes, 1998). As unidades taxonômicas operacionais (UTO) representam as sequências ou os organismos a partir dos quais os dados são provenientes. Os nós internos representam ancestrais hipotéticos e o ancestral comum a todas as sequências ou organismos que compõem a árvore é a raiz da árvore filogenética, no caso de árvores enraizadas (Figura 4) (Page & Holmes, 1998).

As árvores filogenéticas são inferidas a partir da análise de caracteres morfológicos ou moleculares, dentre outros. Na análise molecular, o número de evidências é consideravelmente maior, uma vez que cada nucleotídeo ou aminoácido é interpretado como um caractere que evolui de forma independente (Ridley, 2006). Diferentes proteínas ou segmentos de DNA não evoluem na mesma velocidade e isto deve ser considerado ao determinar a metodologia científica aplicada à reconstrução filogenética. Segundo Mark Ridley (2006) podemos fazer uma analogia com os ponteiros de um relógio que giram em diferentes velocidades: uma molécula de evolução rápida terá “girado” o ponteiro diversas vezes durante o tempo evolutivo. Quando diversas mudanças no mesmo sítio tornam-se comuns, perde-se a informação filogenética. Ou seja, é como tentar utilizar um cronômetro que tem apenas o ponteiro dos segundos para comparar a duração de uma mesma aula, quando ministrada por diferentes professores (Ridley, 2006). Portanto, a velocidade com que as sequências moleculares evoluem deve ser avaliada ao escolher as sequências que serão utilizadas na reconstrução filogenética, de forma que estas estejam de acordo com as perguntas a serem respondidas.

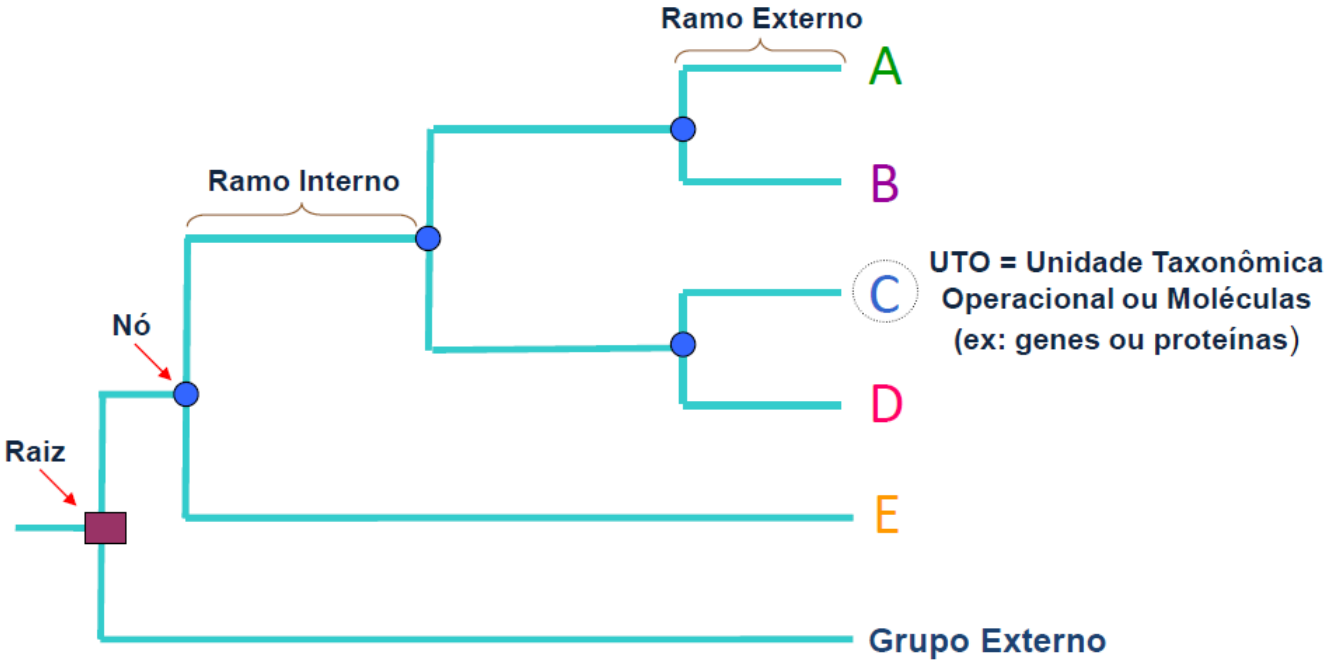


Figura 4. Partições da árvore filogenética. Nesta árvore hipotética as letras A, B, C, D e E representam as sequências proteicas a partir das quais a árvore foi reconstruída. A árvore representa a história evolutiva deste grupo de moléculas e consiste de nós conectados por ramos. Os nós internos representam ancestrais hipotéticos e o ancestral comum a todas as sequências ou organismos que compõem a árvore é a raiz da árvore filogenética. O enraizamento da árvore é realizado a partir do grupo externo.

Há décadas, a filogenética molecular tem sido uma das áreas de pesquisa mais ativas em Biologia Evolutiva. As análises têm início com o alinhamento das sequências, seguido do método de inferência de árvores filogenéticas. Ao comparar sequências moleculares, antes de inferir o número de alterações nestes dados é preciso garantir que os sítios comparados são os mesmos em todas as amostras, motivo pelo qual a probabilidade da hipótese filogenética representar a realidade está diretamente relacionada à qualidade do alinhamento. Portanto, sítios (nucleotídeos ou resíduos de aminoácidos) mal alinhados ou a escolha de sequências não apropriadas podem conduzir a erros que invalidarão a árvore filogenética inferida (Frickey & Lupas, 2004).

Ao longo de várias décadas, diferentes métodos foram descritos para converter a informação armazenada em sequências nucleotídicas ou proteicas em árvores evolutivas. Estes métodos podem ser classificados em dois grandes grupos: os baseados em distância e os baseados em caracteres (Schneider, 2007; Nahum & Pereira, 2008).

Os métodos baseados em distância como o de “agrupamento de vizinhos” (do inglês, *neighbor-joining*, NJ) (Saitou & Nei, 1987; Studier & Keppler, 1988) comparam as diferenças entre todos os pares das sequências alinhadas e em seguida convertem as diferenças em uma matriz de distâncias que representa a distância evolutiva entre cada par. Desta forma, a matriz representa o número ou a proporção de diferenças observadas, ou o número esperado de substituições por sítio sob as premissas de um modelo de evolução molecular (Nahum & Pereira, 2008). As desvantagens dos métodos baseados em distância incluem a perda de informações quando caracteres discretos são convertidos em uma matriz de distância e o fato das estimativas confiáveis de distância não serem obtidas para sequências muito divergentes (Nahum & Pereira, 2008; Holder & Lewis, 2003).

Os métodos baseados em caracteres como máxima verossimilhança (do inglês, *maximum likelihood*, ML) (Felsenstein, 1981) analisam cada caractere individualmente. Este método permite a incorporação de modelos evolutivos complexos na análise produzindo árvores filogenéticas mais acuradas (Holder & Lewis, 2003; Nahum & Pereira, 2008). Entretanto, este método demanda mais tempo e capacidade de processamento computacional.

A utilização de modelos evolutivos na reconstrução de árvores filogenéticas permite incorporar suposições sobre o processo evolutivo que originou os dados observados (Kelchner & Thomas, 2007). A fixação de qualquer mutação depende, dentre outros fatores, da pressão seletiva à qual o gene ou produto gênico estão sujeitos. Regiões codificantes e não codificantes possuem diferentes taxas de substituição, bem como cada posição no códon está sujeita a diferentes pressões seletivas. Por este motivo, a incorporação de heterogeneidade nas taxas evolutivas entre os sítios permite que os modelos se ajustem melhor aos dados produzindo árvores filogenéticas mais realistas (Liò & Goldman, 1998).

Vários modelos evolutivos podem ser testados na reconstrução de árvores filogenéticas, dentre eles: JTT (Jones *et al*., 1992), WAG (Whelan & Goldman, 2001), BLOSUM62 (Henikoff & Henikoff, 1992) e VT (Muller & Vingron, 2000). Para as proteínas codificadas pelo genoma mitocondrial, existem modelos evolutivos específicos como o MtREV (Adachi & Hasegawa, 1996). A seleção do modelo que melhor se adéqua aos dados é de extrema importância para a acurácia da análise filogenética (Posada & Buckley, 2004). Dentre os métodos de seleção disponíveis, o critério de informação de Akaike (AIC) (Akaike, 1973) tem se destacado (Posada & Buckley, 2004). Este critério não testa os modelos no sentido de testar de hipóteses, mas sim avalia os modelos testados. O valor de AIC é derivado dos valores de máxima verossimilhança estimados, penalisando o número de parâmetros analisados. Assim, são atribuídos valores para a qualidade de ajuste de cada modelo sendo, portanto, uma ferramenta para selecionar um modelo evolutivo dentre os demais (Kelchner & Thomas, 2007).

1.2.1. Reconstrução de filomas

As análises filogenéticas de sequências moleculares têm numerosas aplicações como o estabelecimento acurado de relações de ortologia e paralogia, a predição funcional de proteínas não caracterizadas experimentalmente ou a detecção de eventos evolutivos (Huerta-Cepas *et al.*, 2010b).

Nos últimos anos, o crescente número de genomas completamente sequenciados aliados ao desenvolvimento de poder computacional e novas ferramentas de Bioinformática, abriram caminho para as análises comparativas envolvendo diversos grupos taxonômicos (Medina, 2005; Huerta-Cepas *et al.*, 2007), incluindo a reconstrução da história evolutiva de todos os genes codificados em um organismo, i.e. o filoma (Sicheritz-Ponten & Andersson, 2001).

A disponibilização de informações resultantes da reconstrução de filomas provê à comunidade científica uma ampla visão da evolução dos genomas sob a perspectiva de todos os componentes do repertório gênico ou protéico de certa espécie, além de propiciar acesso ao conhecimento em diferentes níveis de divergência da árvore da vida (Medina, 2005; Huerta-Cepas *et al.*, 2007). Estas análises, que integram todos os elementos de um sistema biológico, ao invés de estudar cada componente de forma isolada, deram origem a uma nova abordagem chamada Biologia de Sistemas (Hood, 2003). Esta abordagem possui especial relevância no contexto de organismos recém sequenciados, uma vez que através da genômica ou proteômica comparativa permite realizar a transferência de anotação funcional baseada em relações de ortologia (Huerta-Cepas *et al.*, 2010b). Assim, a correlação entre Teoria Evolutiva e Biologia de Sistemas surge para solucionar duas das questões fundamentais em Biologia: a função dos sistemas biológicos e a compreensão da evolução da biodiversidade (Medina, 2005).

1.2.2. Anotação funcional de genes e seus produtos

A predição funcional de genes e produtos gênicos é uma etapa chave em qualquer projeto de sequenciamento genômico. Entretanto, embora o conhecimento da função gênica seja de extrema importância para a compreensão dos processos biológicos, cerca de 30-35% dos genes, na maioria dos organismos, permanece sem caracterização experimental ou predita e grande parte das demais sequências codificantes possuem apenas uma

anotação funcional muito genérica (Galperin & Koonin, 2000). Por este motivo, pesquisas têm sido desenvolvidas com o objetivo de anotar automaticamente as funções proteicas em resposta à alta taxa de geração de sequências em comparação com a taxa de caracterização experimental (Galperin & Koonin, 2000; Gabaldón & Huynen, 2004; Sjölander, 2004; Hawkins & Kihara, 2007; Nahum & Pereira, 2008; Gabaldón, 2008a; Jiang, 2008). Neste sentido, duas abordagens têm sido utilizadas com maior freqüência: uma baseada em similaridade de sequências e outra baseada em análise filogenética.

A anotação funcional dos produtos gênicos de *S. mansoni* (Berriman *et al.*, 2009) foi realizada através de buscas baseadas em similaridade de sequências utilizando ferramentas do pacote BLAST – *Basic Local Alignment Search Tool* (Altschul *et al.*, 1990). Este método é rápido, simples e pode ser facilmente automatizado para a análise de milhares de genes, motivo pelo qual tem sido utilizado com frequência para a predição funcional de produtos codificados por genomas recém sequenciados. Nestes casos, os *top hits* que retornam do banco de dados são interpretados como estreitamente relacionados ao gene ou proteína que motivou a busca. Entretanto, várias limitações estão associadas a este método como, por exemplo, o fato dos resultados obtidos através de alinhamento local não refletirem a função proteica totalmente. Em conjunto, a prática de predição funcional baseada em similaridade de sequências tem gerado erros sistemáticos (Brenner, 1999; Galperin & Koonin, 1998; Koski & Golding, 2001; Brown & Sjolander, 2006; Nahum & Pereira, 2008).

Na tentativa de aumentar a acurácia da predição funcional em larga escala, cientistas têm associado métodos filogenéticos a esta análise, tirando proveito do poder que uma perspectiva evolutiva fornece às análises biológicas comparativas (Eisen, 1998; Eisen & Wu 2002). Certo das vantagens proporcionadas por esta análise composta, Eisen (1998) sugeriu que as análises evolutivas e genômicas devem ser combinadas em uma única abordagem, denominada como *Filogenômica*. Assim, há aproximadamente 12 anos, a filogenômica foi proposta como a interseção entre a genômica e os estudos filogenéticos. Desde então, este novo campo de atuação tem provido informações importantes no que diz respeito à predição funcional dos genes, bem como sobre a história evolutiva dos genomas e seus respectivos organismos, tendo se mostrado uma alternativa eficaz para otimizar o processo de predição funcional e evitar a propagação de erros associados à predição baseada apenas em similaridade de sequências (Eisen, 1998; Frickey & Lupas, 2004; Brown & Sjolander, 2006; Nahum & Pereira, 2008; Huerta-Cepas *et al.*, 2010b).

Em virtude do exposto e da ausência de anotação funcional para mais de 42% do proteoma predito de *S. mansoni* (Berriman *et al.*, 2009, Zerlotini *et al*., 2009), o presente trabalho tem como motivação principal aprimorar a predição funcional do respectivo proteoma deste parasito, baseado nos potenciais da análise comparativa sob uma perspectiva evolutiva.

2. OBJETIVOS

2.1. Objetivo geral:

2.1.1. Aprimorar a anotação funcional do genoma de *S. mansoni* utilizando a abordagem filogenômica para a predição de ortólogos.

2.2. Objetivos específicos

2.2.1. Reconstruir e analisar o filoma de *Schistosoma mansoni* a partir do proteoma predito deste parasito e de seus potenciais homólogos em 16 outros organismos.

2.2.2. Construir um banco de dados relacional que integre dados do SchistoDB, PhylomeDB e resultados de nossas análises.

2.2.3. Identificar e remover dados do proteoma de *S. mansoni* que sejam potencialmente redundantes.

2.2.4. Analisar as relações evolutivas de *S. mansoni* com outros metazoários e identificar genes e/ou grupos de genes que foram duplicados nesta espécie.

3. MATERIAIS E MÉTODOS

3.1. Seleção dos proteomas preditos

O filoma reconstruído neste trabalho é derivado do conjunto de proteínas codificadas pelo genoma completamente sequenciado de 17 organismos obtidos no *Ensembl*, *Intergr8*, *JGI Genome Projects*, *SchistoDB*, *Broad Institute Ustilago maydis* Database, *SilkDB* e *NCBI Genome Database* (Quadro 1). As proteínas mitocondriais foram recuperadas separadamente a partir do *GOBASE* (O'Brien *et al*., 2009) e do banco de dados de organelas eucariotas do *NCBI* (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome).

Quadro 1*:* Base de dados utilizada na reconstrução do filoma de *S. mansoni*

Nome científico	Código[1]	TaxID[2]	Proteínas[3]	Fonte[4]	Referência[5]
Arabidopsis thaliana	Ath	3702	35.552	http://www.ebi.ac.uk/integr8	*Arabidopsis* Genome Initiative, 2000
Ustilago maydis	Uma	5270	6.548	http://www.broadinstitute.org/annotation/genome/ustilago_maydis/	*Ustilago maydis* Sequencing Project. Broad Institute of MIT and Harvard (http://www.broad.mit.edu)
Saccharomyces cerevisiae	Sce	4932	6.600	http://fungi.ensembl.org/Saccharomyces_cerevisiae	Cherry *et al.*, 1997; Liti *et al*., 2009
Neurospora crassa	Ncr	5141	10.645	http://www.broadinstitute.org/annotation/genome/neurospora	Galagan *et al.*, 2003
Nematostella vectensis	Nem	45351	27.175	http://genome.jgi-psf.org/Nemve1/Nemve1.download.ftp.html	Putnam *et al.*, 2007
Caenorhabditis elegans	Cel	6239	19.574	http://www.ensembl.org/Caenorhabditis_elegans	The *C. elegans* Sequencing Consortium, 1998
Caenorhabditis briggsae	Cbr	6238	13.192	http://metazoa.ensembl.org/Caenorhabditis_briggsae	Stein *et al.*, 2003
Schistosoma mansoni	Sch	6183	13.285	http://schistodb.net	Berriman *et al.*, 2009
Drosophila melanogaster	Dme	7227	13.734	http://metazoa.ensembl.org/Drosophila_melanogaster	Adams *et al.*, 2000;
Anopheles gambiae	Aga	7165	14.244	http://metazoa.ensembl.org/Anopheles_gambiae	Holt *et al.*, 2002; Sharakhova *et al*., 2007
Bombyx mori	Bom	7091	14.623	http://silkworm.genomics.org.cn	Mita *et al.*, 2004
Strongylocentrotus purpuratus	Str	7668	42.420	http://www.ncbi.nlm.nih.gov/projects/genome/guide/sea_urchin/	Sea Urchin Genome Sequencing Consortium *et al.*, 2006
Ciona intestinalis	Cin	7719	10.978	http://www.ensembl.org/Ciona_intestinalis	Dehal *et al.*, 2002
Branchiostoma floridae	Bra	7739	50.817	http://genome.jgi-psf.org/Brafl1/Brafl1.home.html	Putnam *et al.*, 2008
Danio rerio	Dre	7955	21.409	http://www.ensembl.org/Danio_rerio	Almeida-King *et al.*, 2009.
Mus musculus	Mms	10090	25.252	http://www.ensembl.org/Mus_musculus	Church *et al*., 2009
Homo sapiens	Hsa	9606	21.926	http://www.ensembl.org/Homo_sapiens	International Human Genome Sequencing Consortium, 2004

1 – Código da espécie analisada. 2 – Identificador taxonômico no NCBI (taxID). 3 – Total de proteínas analisadas em cada espécie. 4 – Bancos de dados dos proteomas preditos. 5 – Referências bibliográficas.

O conjunto de espécies selecionadas corresponde principalmente a sequências de metazoários (animais heterotróficos, móveis e multicelulares), incluindo 8 invertebrados (*Nematostella vectensis*, *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Schistosoma mansoni*, *Drosophila melanogaster*, *Anopheles gambiae*, *Bombyx mori* e *Strongylocentrotus purpuratus*), 1 tunicado (*Ciona intestinalis*), 1 cefalocordado (*Branchiostoma floridae*) e 3 vertebrados (*Danio rerio*, *Mus musculus* e *Homo sapiens*) (Quadro 1). Esta amostragem de espécies torna este banco de dados bastante adequado para o estudo da evolução de famílias de proteínas entre os metazoários.

3.2. Reconstrução do filoma de *S. mansoni*

Para reconstruir o filoma de *S. mansoni* foi utilizado um *pipeline* automático (Figura 5), adaptado a partir do descrito para a reconstrução do filoma humano (Huerta-Cepas *et al.*, 2007). Um *pipeline* é um conjunto de passos ou instruções computacionais a serem executadas para certa finalidade, podendo ser automatizada, por exemplo, em análises de Bioinformática.

Para cada proteína codificada pelo genoma de *S. mansoni*, foi realizada uma busca por similaridade (Altschul, 1990) contra o banco de dados gerado, para selecionar um conjunto de proteínas com índice de similaridade significativo (E-value $<10^{-3}$). Somente sequências que alinharam com cobertura mínima de 33% em relação à sequência proteica em análise foram selecionadas, com um limite máximo de 150 sequências por busca (ordenadas de forma crescente pelo valor do E-value). Os conjuntos de proteínas recuperados pela busca por similaridade foram alinhados utilizando os parâmetros padrão do MUSCLE, versão 3.6 (Edgar, 2004), um programa que realiza o alinhamento global de múltiplas sequências. Para filtrar colunas (sítios) do alinhamento ricas em lacunas (*gaps)*, aquelas contendo *gaps* em mais de 10% das sequências foram eliminadas utilizando o programa trimAl (Capella-Gutierrez *et al.*, 2009) com limiar para *gaps* em cada coluna igual a 0.9 e com limite de cobertura de, no mínimo, 33%. A aplicação de filtro nos alinhamentos "brutos" (*whole alignment*) produz alinhamentos mais "limpos" (*trimmed alignment*) mantendo apenas trechos de sequências compartilhadas pela maioria dos membros selecionados na busca por similaridade.

A partir dos alinhamentos filtrados foram obtidas árvores filogenéticas utilizando duas abordagens: uma baseada em distância - *Neighbor Joining* (NJ) (Saitou & Nei, 1987; Studier & Keppler, 1988) e outra baseada em caracteres - *Maximum Likelihood* (ML) (Felsenstein, 1981). As árvores filogenéticas reconstruídas através da abordagem NJ foram obtidas utilizando uma matriz de distância, como implementado no BioNJ (Gascuel, 1997). Em

seguida, estas foram utilizadas como "semente" para a análise por ML utilizando o algoritmo PhyML (Guindon & Gascuel, 2003). O princípio desta análise é começar a partir de uma árvore construída por um algoritmo rápido e melhorá-la. Na análise por ML, para cada árvore filogenética foram testados quatro modelos evolutivos: JTT (Jones *et al*., 1992), WAG (Whelan & Goldman, 2001), BLOSUM62 (Henikoff & Henikoff, 1992) e VT (Muller and Vingron, 2000). Para as proteínas codificadas pelo genoma mitocondrial o *pipeline* foi aplicado separadamente, e além dos quatro modelos mencionados, acrescentamos o modelo MtREV (Adachi & Hasegawa, 1996).

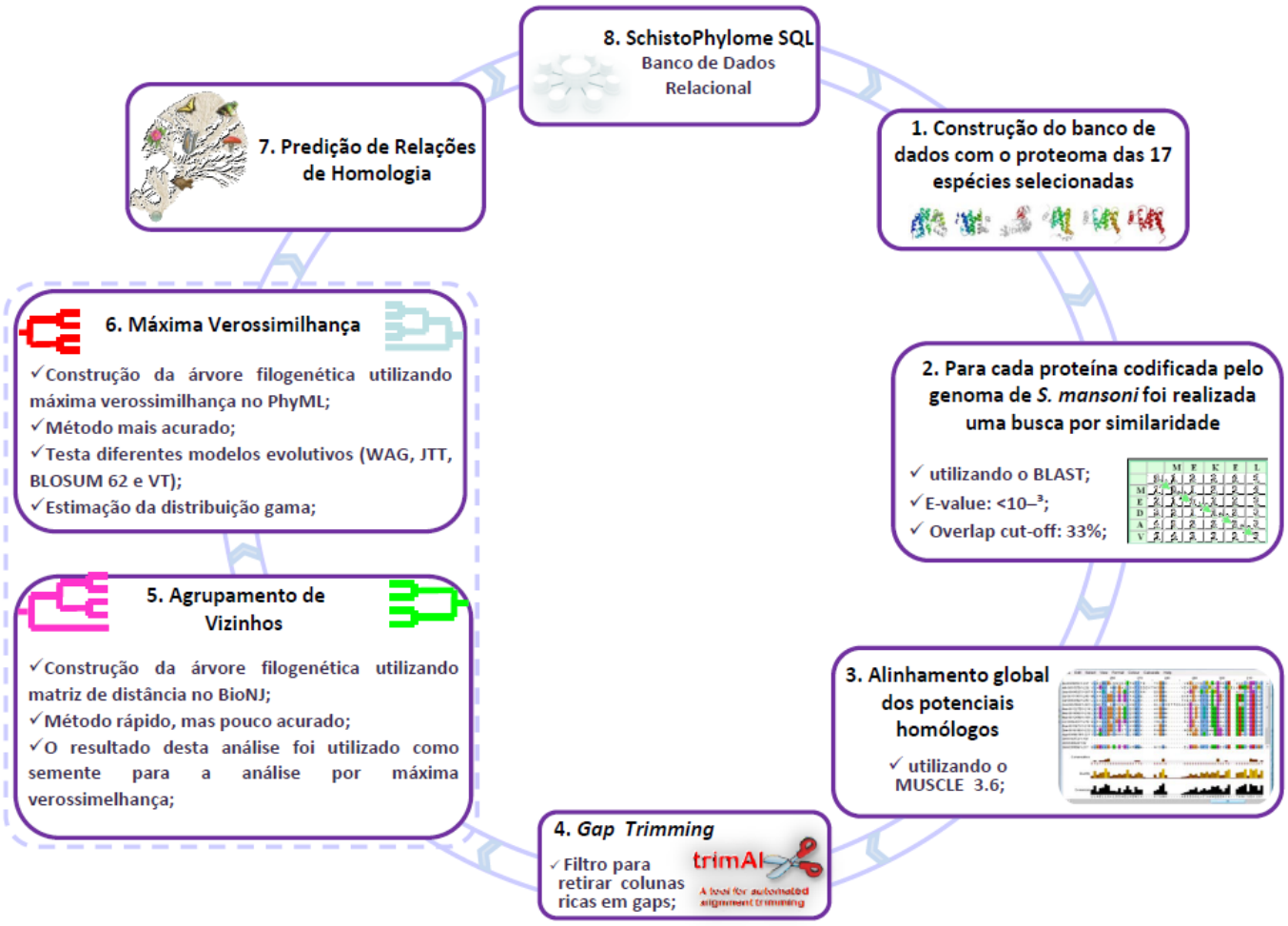


Figura 5: Representação esquemática do *pipeline* utilizado na reconstrução do filoma de *S. mansoni*. Cada sequência proteica codificada pelo genoma de *S. mansoni* é comparada contra o banco de dados contendo o proteoma predito das 16 espécies incluídas na análise para selecionar possíveis proteínas homólogas. O grupo de sequências selecionadas é alinhado e, em seguida, filtrado para a remoção de colunas ricas em *gaps*. O alinhamento refinado é utilizado na reconstrução da árvore filogenética pelo método de distância (Agrupamento de Vizinhos), a qual é então utilizada como árvore "semente" para a análise por máxima verossimilhança. Nesta última, são testados quatro modelos evolutivos (WAG, JTT, BLOSUM 62 e VT) (cinco, no caso de proteínas codificadas pelo genoma mitocondrial onde testa-se também o MtREV). (Adaptado de Huerta-Cepas *et al.*, 2007).

Os modelos evolutivos foram utilizados para fazer predições quanto ao processo de substituição nas sequências moleculares, ou seja, um modelo de substituição de sequências descreve em termos probabilísticos o processo através do qual uma sequência de caracteres (nucleotídeos ou aminoácidos) evoluiu ao longo do tempo (Liò & Goldman, 1998; Kelchner & Thomas, 2007). Uma vez que nem todos os sítios evoluem na mesma taxa, em função de alguns sítios estarem sobre forte pressão seletiva, aplicamos em todos os casos um modelo de distribuição gama discreta, com quatro categorias além de posições invariantes. O parâmetro gama e a fração de posições invariantes foram estimados a partir dos dados.

Para selecionar o modelo evolutivo que melhor se adequou aos dados utilizamos o critério de informação de Akaike (AIC) (Akaike, 1973). Valores de apoio para as diferentes partições das árvores filogenéticas foram calculados pelo teste aLRT (do inglês, *approximate likelihood ratio test*), como implementado no PhyML (Guindon & Gascuel, 2003; Anisimova & Gascuel, 2006), utilizando a opção *Chi2-Based*. Esta abordagem avalia cada ramo da árvore calculando se o ramo em análise possui ganho de verossimilhança significativo em comparação com a hipótese nula que determina o colapso deste ramo na árvore filogenética (Anisimova & Gascuel, 2006).

3.3. Construção do banco de dados relacional

Banco de dados relacional é um conceito abstrato que define maneiras de armazenar, manipular e recuperar dados estruturados na forma de entidades (tabelas) (Ferreira & Takai, 2008). Estes dados são registros de qualquer natureza, inclusive dados biológicos, como por exemplo, identificadores de genes, sequências nucleotídicas ou proteicas, alinhamentos ou árvores filogenéticas. As tabelas são formadas por linhas e colunas. As colunas são indexadas possibilitando melhor desempenho nos processos de busca e ordenação dos dados. O estabelecimento de relações entre as tabelas é realizado através de uma chave primária - um identificador único constituído por uma ou mais colunas.

Para construir um banco de dados, é necessário um Sistema Gestor de Base de Dados (SGBD) - um conjunto de programas responsáveis pelo gerenciamento da base. No presente trabalho, utilizamos o MySQL, versão 5.1.49 (http://www.mysql.com/downloads/mysql/) como SGBD que executa comandos em Linguagem de Consulta Estruturada (do inglês, *Structured Query Language* - SQL).

Em função da necessidade de extrair diversas informações a partir de grandes conjuntos de dados de fontes diferentes, utilizando as ferramentas acima mencionadas,

construímos o banco de dados relacional SchistoPhylomeSQL (Figura 8). Este banco integra informações do proteoma de *S. mansoni* disponíveis no *SchistoDB*, versão 2.0 (Zerlotini *et al.*, 2009) com dados do proteoma deste parasita obtidos no UniProt (The UniProt Consortium, 2010), além de dados do filoma depositados no *PhylomeDB* (Huerta-Cepas *et al.*, 2008). O acesso ao banco de dados foi realizado através do DbVisualizer, versão 7.0.5 (http://www.dbvis.com), uma ferramenta intuitiva que permite desenvolver e acessar diferentes SGDBs em vários sistemas operacionais.

O banco de dados SchistoPhylomeSQL (Figura 8) foi o principal recurso para a mineração de dados do filoma de *S. mansoni* descrita a seguir.

3.4. Mineração dos dados

As análises comparativas desenvolvidas neste trabalho foram realizadas com base no proteoma predito de *S. mansoni,* disponível no banco de dados *SchistoDB* 2.0 (Zerlotini *et al.*, 2009), em conjunto com os demais 16 proteomas já mencionados (Quadro 1). A mineração dos dados selecionados para integrar a análise bem como dos dados obtidos como resultados foram realizadas via consultas em SQL no banco de dados relacional SchistoPhylomeSQL.

Para acessar o banco de dados, inicialmente é necessário abrir a interface interativa do DbVisualizer e em seguida estabelecer conexão com os dados do SchistoPhylomeSQL através da identificação do sistema gestor de base de dados, do servidor, do banco de dados e do usuário. Uma vez conectados, todas as tabelas podem ser facilmente acessadas e as consultas em SQL efetuadas através do campo *SQL Commander*. Exemplos de consultas em SQL utilizadas na mineração do filoma podem ser obtidos no Quadro 2.

O estabelecimento de relações entre as tabelas foi realizado através dos identificadores originais das sequências no *SchistoDB* e/ou no *PhylomeDB*. Os resultados das consultas deram origem às estatísticas e gráficos do filoma de *S. mansoni*.

Quadro 2*:* Exemplos de consultas em SQL utilizadas na mineração do filoma de *S. mansoni.*

Descrição[1]	Total[2]	Consulta em SQL[3]
Transcritos analisados	13.285	select count(*) from idconversion where originalid like 'Smp_%' or originalid like 'NC%'
Alinhamentos trimados	8.818	select count(*) from alignment where type like 'trimmed'
Árvores filogenéticas	8.818	select count(*) proteinid from SchistoBestModel
Proteínas hipotéticas ou expressas sem termos Gene Ontology	5.631	select count(*) from SchistoAnnotation a where (a.product like '%hypothetic%' or a.product like '%express%') and a.go not like '%GO%'
Árvores para proteínas hipotéticas ou expressas sem termos Gene Ontology	2.534	select count(*) from SchistoBestModel bm,SchistoAnnotation a where bm.proteinid=a.proteinid and (lower(a.product) like '%hypothetic%' or lower(a.product) like '%express%') and (lower(a.go) not like '%GO%')
Árvores filogenéticas que possuem apenas identificadores de *S. mansoni* e cujo identificador *seed* possui como produto a palavra *hypothetical* ou *expressed protein*	966	select count(*) from SchistoBestModel bm, SchistoAnnotation a where a.proteinid=bm.proteinid and (lower(a.product) like '%ypothetic%' or lower(a.product) like '%xpress%') and bm.newick not like '%Dre%' and bm.newick not like '%Bra%' and bm.newick not like '%Aga%' and bm.newick not like '%Mms%' and bm.newick not like '%Ath%' and bm.newick not like '%Nem%' and bm.newick not like '%Sce%' and bm.newick not like '%Bom%' and bm.newick not like '%Cbr%' and bm.newick not like '%Cel%' and bm.newick not like '%Cin%' and bm.newick not like '%Hsa%' and bm.newick not like '%Uma%' and bm.newick not like '%Dme%' and bm.newick not like '%Str%' and bm.newick not like '%Ncr%'

1 – Descrição do que a consulta objetiva identificar. 2 – Total de identificadores encontrados. 3 – Consulta em SQL utilizada.

3.5. Identificação de produtos de *splicing* alternativo e sequências redundantes

Com o objetivo de filtrar os produtos de *splicing* alternativo (isoformas) presentes no banco de dados *SchistoDB* 2.0 (Zerlotini *et al.*, 2009), utilizando o SchistoPhylomeSQL, recuperamos uma lista contendo todos os identificadores das proteínas e o tamanho das sequências (aa) para os quais existiam mais de uma versão (exemplo, Smp_000130.1, Smp_000130.2, Smp_000130.3 e Smp_000130.4). Diferentes versões são transcritos produto de *splicing* alternativo codificados pelo mesmo gene. Esta lista possuía inicialmente um total de 2.188 identificadores. No caso de múltiplas versões da mesma proteína, os dez primeiros números do identificador são iguais e as diferentes versões são separadas por ponto (.) seguido do número que indica a versão. Por exemplo, Smp_000130.1, Smp_000130.2, Smp_000130.3 e Smp_000130.4 representam quatro diferentes versões codificadas pelo gene Smp_000130. Em seguida, um programa escrito em linguagem Python, versão 2.5.5 (http://www.python.org/download/releases), analisou as sequências

comparando o tamanho entre as versões e recuperando apenas a de maior tamanho em aminoácidos.

Para filtrar a redundância por similaridade, através do ETE (*Enviroment for tree Exploration*) (http://ete.cgenomics.org/) (Huerta-Cepas *et al*., 2010a), conectamos no *PhylomeDB* e selecionamos todas as árvores contendo, no mínimo, duas sequências de *S. mansoni.* Este procedimento nos permitiu recuperar um total 6.118 árvores para as quais buscamos os alinhamentos não filtrados (*whole alignment*). Para identificar a porcentagem de identidade entre as sequências de *S. mansoni* utilizamos o programa trimAl em cada alinhamento (Capella-Gutierrez *et al.*, 2009). Em seguida, um programa escrito em linguagem Python identificou as sequências com porcentagem de identidade igual ou maior a 98% e selecionou de forma aleatória apenas uma delas, descartando as demais.

3.6. Análise filogenética das proteínas mitocondriais

O número reduzido de genes mitocondriais e o variado grau de conservação entre eles torna o genoma mitocondrial adequado tanto para o estudo intra-específico quanto para o estudo inter-específico. Como todas as espécies incluídas na reconstrução do filoma de *S. mansoni* possuem o genoma mitocondrial completamente sequenciado e anotado, foi realizado um levantamento prévio que verificou a presença de sequências homólogas às proteínas mitocondriais de *S. mansoni* em todos os organismos. Por este motivo e em função de proteínas como Citocromo *b* (CYTB) e Citocromo Oxidase (COX) serem altamente conservadas (Saribas *et al*., 1998; Zhao *et al*., 2008), utilizamos as árvores filogenéticas reconstruídas para as proteínas CYTB e COXI como grupo controle para a metodologia utilizada na seleção de possíveis ortólogos durante a reconstrução do filoma de *S. mansoni*.

A visualização das árvores filogenéticas no formato gráfico foi realizada utilizando o Figtree (Rambaut, 2010). Para validação da metodologia aplicada, deve ser possível observar nas árvores mencionadas a sequência codificada pelo genoma mitocondrial de *S. mansoni* e seus homólogos em cada um dos demais16 organismos.

3.7. Detecção de relações de homologia

Com o objetivo de gerar um catálogo completo com a predição de homólogos entre os genes de *S. mansoni* e os demais organismos incluídos na análise (Quadro 1), examinamos o filoma de *S. mansoni* com um algoritmo que realiza a predição de ortologia e

paralogia (Huerta-Cepas *et al.*, 2007) através de *scripts* e programas gerados em linguagem Python para exploração de árvores filogenéticas (http://ete.cgenomics.org/) (Huerta-Cepas *et al*., 2010a). Em síntese, este algoritmo examina cada árvore filogenética e utiliza a distribuição das espécies entre grupos irmãos para detectar e marcar eventos de duplicação gênica e especiação (Figura 6). Para cada proteína "semente", o algoritmo segue pelos ramos e nós que a conectam com as demais proteínas presentes na árvore filogenética e analisando o evento evolutivo predito em cada nó o algoritmo estabelece relações de paralogia e ortologia. Ao longo deste trabalho nos referimos às proteínas de *S. mansoni* como proteínas "semente" em função de todas as análises terem sido baseadas nestas sequências.

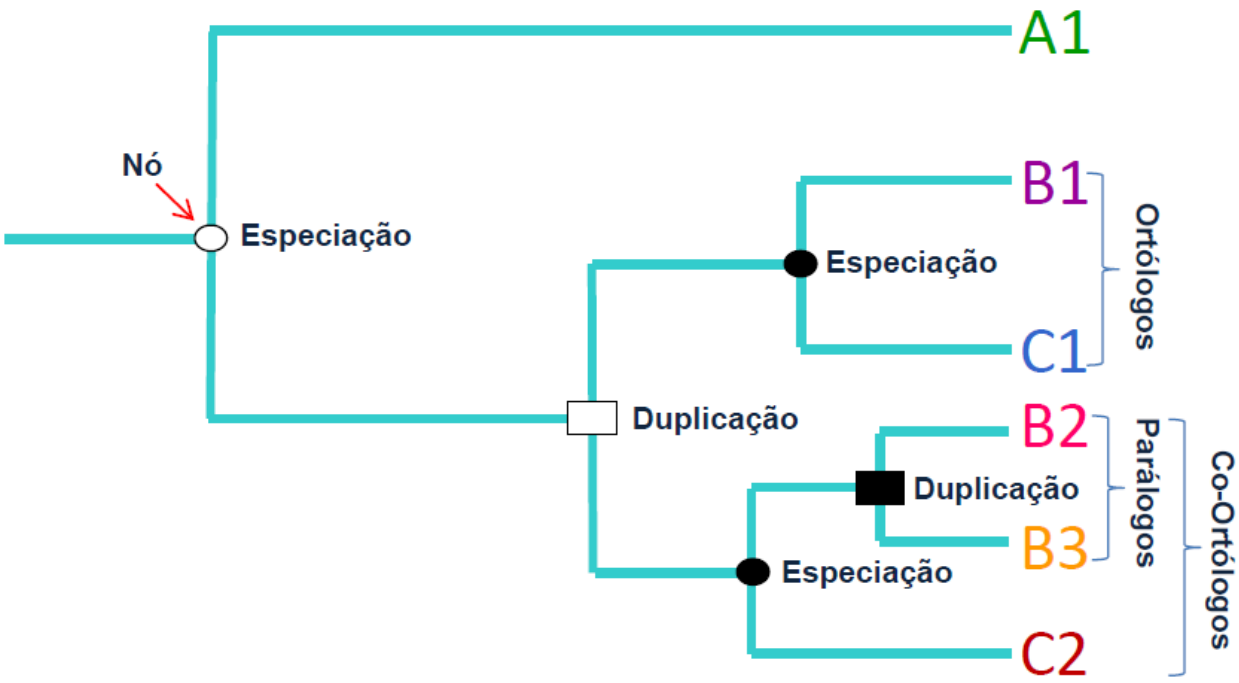


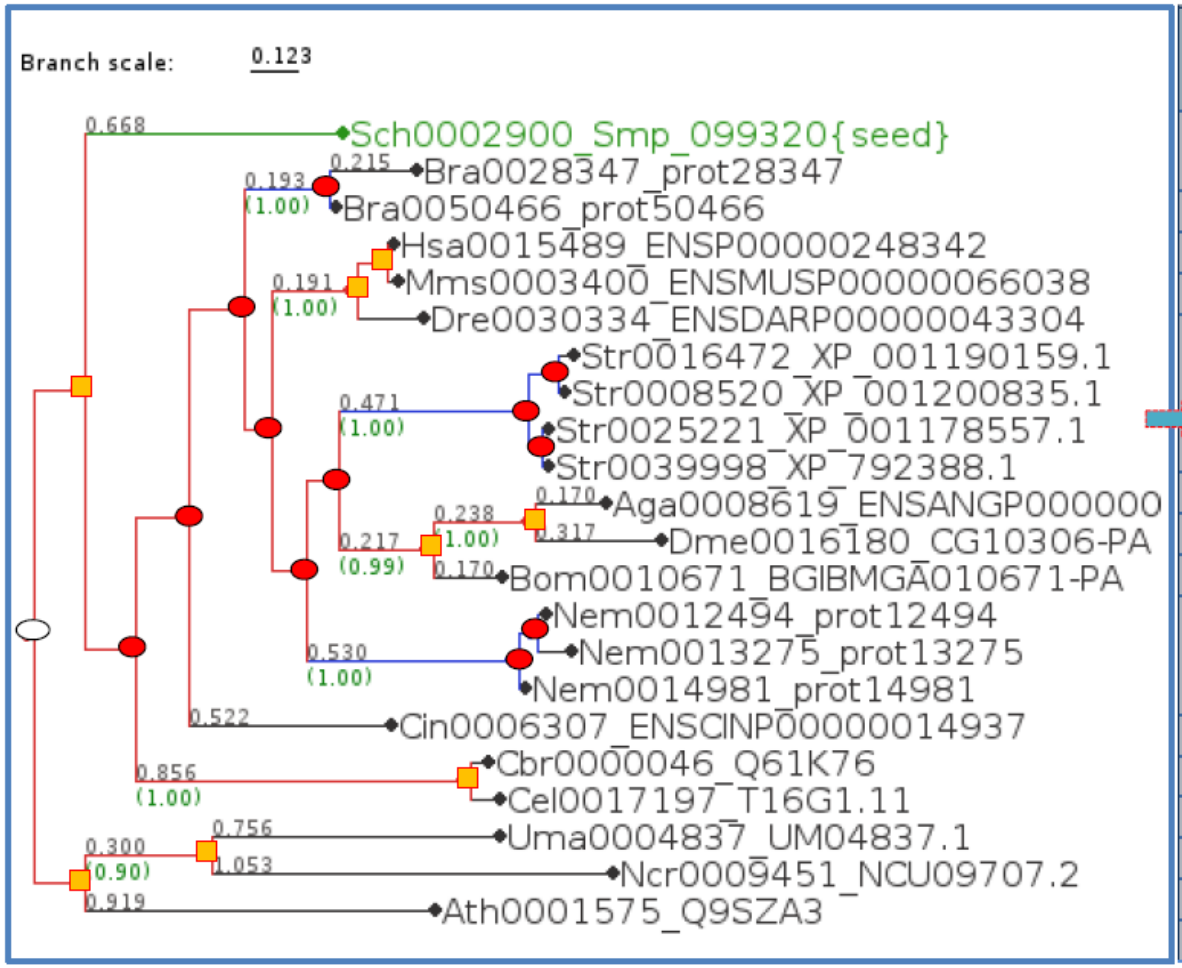
Figura 6: Relações de ortologia e paralogia em uma família de proteínas. A árvore filogenética de uma família de proteínas hipotética contem um total de seis membros: um membro da espécie A (A1), três da espécie B (B1, B2 e B3) e dois da espécie C (C1, C2). Eventos de especiação (círculos) e duplicação gênica (quadrados) estão indicados. Um primeiro evento de duplicação (quadrado branco) ocorreu antes da especiação B-C (círculo preto), enquanto uma duplicação posterior (quadrado preto) ocorreu na linhagem B dando origem à relação de paralogia entre B2 e B3. Neste cenário, a única proteína presente em A (A1) é ortóloga a todas as demais e vice-versa. Quando comparamos membros de B e C: C1 é ortólogo de B1, enquanto C2 possui dois co-ortólogos em B (B2 e B3) (Adaptado de Gabaldón, 2005).

3.8. Inferência da função gênica por métodos filogenômicos

A fim de produzir um conjunto de predições funcionais com alta confiabilidade para as proteínas codificadas pelo genoma de *S. mansoni*, a partir dos dados de relações evolutivas previamente obtidos e filtrados, recuperamos um subconjunto de ortologias com relações do tipo “um-pra-um”, “um-pra-muitos”, “muitos-pra-um” e “muitos-pra-muitos” entre *S. mansoni* e cada um dos demais 16 organismos. Os critérios utilizados para o agrupamento em subconjuntos são:

- ✓ um-pra-um: quando a proteína “semente” não possui proteínas parálogas em *S. mansoni* e existe uma única proteína ortóloga na outra espécie analisada (Remm *et al.*, 2001);
- ✓ um-pra-muitos: quando a proteína “semente” não possui proteínas parálogas em *S. mansoni* e existe mais de uma proteína ortóloga na outra espécie analisada (Remm *et al.*, 2001);
- ✓ muitos-pra-um: quando a proteína “semente” possui proteínas parálogas em *S. mansoni* e existe uma única proteína ortóloga na outra espécie analisada (Remm *et al.*, 2001);
- ✓ muitos-pra-muitos: quando a proteína “semente” possui proteínas parálogas em *S. mansoni* e existe mais de uma proteína ortóloga na outra espécie analisada (Remm *et al.*, 2001);

Exemplos de subconjuntos agrupados de acordo com as relações de homologia podem ser visualizados na Figura 7 e nos Quadros 3 e 9.



"Semente"	Ortólogos	Relações de Homologia
Sch0002900	Bra0050466 Bra0028347	um-pra-muitos
Sch0002900	Dre0030334	um-pra-um
Sch0002900	Hsa0015489	um-pra-um
Sch0002900	Mms0003400	um-pra-um
Sch0002900	Str0025221 Str0039998 Str0016472 Str0008520	um-pra-muitos
Sch0002900	Bom0010671	um-pra-um
Sch0002900	Aga0008619	um-pra-um
Sch0002900	Dme0016180	um-pra-um
Sch0002900	Nem0012494 Nem0013275 Nem0014981	um-pra-muitos
Sch0002900	Cin0006307	um-pra-um
Sch0002900	Cbr0000046	um-pra-um
Sch0002900	Cel0017197	um-pra-um
Sch0002900	Ath0001575	um-pra-um
Sch0002900	Uma0004837	um-pra-um
Sch0002900	Ncr0009451	um-pra-um

Figura 7: Predição de eventos evolutivos e relações de homologia das proteínas presentes na árvore filogenética reconstruída para a proteína Smp_099320. A análise tem início no ramo externo que conecta a proteína "semente" (*seed*) à árvore filogenética. Em seguida, o programa analisa todos os nós que conectam esta proteína às demais sequências proteicas. De acordo com a distribuição das espécies nas partições descendentes de cada nó, o programa prediz os eventos evolutivos e estabelece relações de homologia entre as proteínas. Após identificar os ortólogos e/ou páralogos e classificar as relações de homologia, esta informação é armazenada em um arquivo texto.

Quadro 3: Exemplo de relações evolutivas entre proteínas de *S. mansoni* e cada um dos demais 16 organismos incluídos na análise.

Proteína 'Semente'[1]	Parálogos em *S. mansoni*[2]	Homólogos[3]	Relações de Homologia[4]
Sch0002900	-	Hsa0015489	um-pra-um
Sch0002900	-	Dme0016180	um-pra-um
Sch0002900	-	Ath0001575	um-pra-um
Sch0002900	-	Ncr0009451	um-pra-um
Sch0002900	-	Cel0017197	um-pra-um
Sch0002900	-	Cin0006307	um-pra-um
Sch0002900	-	Aga0008619	um-pra-um
Sch0002900	-	Cbr0000046	um-pra-um
Sch0002900	-	Nem0012494, Nem0013275, Nem0014981	um-pra-muitos
Sch0002900	-	Uma0004837	um-pra-um
Sch0002900	-	Mms0003400	um-pra-um
Sch0002900	-	Str0025221, Str0039998,Str0016472, Str0008520	um-pra-muitos
Sch0002900	-	Dre0030334	um-pra-um
Sch0002900	-	Bra0050466,Bra0028347	um-pra-muitos
Sch0002900	-	Bom0010671	um-pra-um
Sch0007902	Sch0012555, Sch0007910	Cel0005193	muitos-pra-um
Sch0007902	Sch0012555, Sch0007910	Cin0016615	muitos-pra-um
Sch0007902	Sch0012555, Sch0007910	Mms0024944, Mms0003391	muitos-pra-muitos
Sch0007902	Sch0012555, Sch0007910	Cbr0007035	muitos-pra-um
Sch0007902	Sch0012555, Sch0007910	Nem0025875	muitos-pra-um
Sch0007902	Sch0012555, Sch0007910	Bom0008139	muitos-pra-um
Sch0007902	Sch0012555, Sch0007910	Str0038612, Str0032349	muitos-pra-muitos
Sch0007902	Sch0012555, Sch0007910	Dre0023109, Dre0020056, Dre0020008	muitos-pra-muitos
Sch0007902	Sch0012555, Sch0007910	Bra0001053, Bra0019084	muitos-pra-muitos

1 – Identificador da proteína "semente" de *S. mansoni* na árvore filogenética. 2 – Parálogos em *S. mansoni* para a proteína "semente"; a ausência de parálogos é representada por um traço (–). 3 – Homólogos nos demais organismos. 4 – Classificação da relação de homologia.

Os dados de anotação funcional conhecidos para os genes das espécies incluídas no filoma de *S. mansoni* foram depositados dos bancos de dados (Quadro 1) e armazenados separadamente em arquivos texto nomeados de acordo com o código de cada espécie, como por exemplo, Nem.txt, Aga.txt e Hsa.txt. A partir do arquivo texto contendo as relações de homologia (e.g. Figura 7), um programa escrito em Python (Quadro 4) foi utilizado para efetuar a transferência da anotação para as proteínas de *S. mansoni*. Em síntese, o programa lê cada linha do arquivo contendo os homólogos preditos para cada

proteína “semente” de *S. mansoni* e ao detectar o identificador de um ortólogo, ele acessa o arquivo contendo os dados de anotação de uma determinada espécie, verifica a disponibilidade de anotação funcional para o identificador em análise e efetua a transferência. Caso exista mais de um ortólogo para a mesma proteína “semente”, o programa ordena os dados de anotação funcional de acordo com a ordenação dos ortólogos. Os arquivos resultantes contendo a informação transferida possuem as seguintes informações: identificador da proteína “semente”, identificador do ortólogo, classificação da relação de homologia (um-pra-um, um-pra-muitos, muitos-pra-um ou muitos-pra-muitos) e termos *Gene Ontology* (GO). A metodologia aplicada gera resultados independentes, uma vez que a transferência é feita a partir de cada ortólogo com função conhecida em cada espécie para a proteína “semente”, o que nos permite comparar as anotações funcionais de ortólogos de diferentes espécies. Por exemplo, a proteína mitocondrial Citocromo Oxidase Subunidade I está presente em todas as demais 16 espécies incluídas no filoma de *S. mansoni*. Esta proteína possui anotação funcional no genoma mitocondrial de todas as 16 espécies. Portanto, o *pipeline* utilizado deve ser capaz de identifcar os 16 ortólogos na busca por similaridade e deve ser possível observar como resultado 16 transferências de anotação funcional para o identificador da proteína Citocromo Oxidase Subunidade I de *S. mansoni*.

Quadro 4: Programa escrito em linguagem Python para efetuar a transferência da anotação funcional das proteínas ortólogas com função conhecida para as proteínas "semente" de *S. mansoni*

```
#!/usr/bin/env python

import sys

#Reads orthology file
orthologs = {}
paralogs = {}
type_orthology = {}
list_species = []
for line in open(sys.argv[1]):
	line = line.strip()
	dades = line.split(" ")
	if "Sch" in dades[1]:
		paralogs[dades[0]] = dades[1]
	else:
		if dades[1][0:3] not in list_species:
			list_species.append(dades[1][0:3])
		if dades[0] not in orthologs:
			orthologs[dades[0]] = {}
		orthologs[dades[0]][dades[1][0:3]] = dades[1]

#Read GO files

annotations = {}
for species in list_species:
	annotations[species] = {}
	for line in open("../GO_annotations/GO/"+species+".txt"): #Change is GO files are in a different location
		line = line.strip()
		if "GO:" in line:
			dades = line.split("\t")
			annotations[species][dades[0]] = dades[2]

for species in list_species:
	file = open("./prediction98_version3/"+species+"_annotations","w") #Change if you want a different output
	for seed in paralogs.keys():
		if "," not in paralogs[seed]:
			if species in orthologs[seed]:
				if "," not in orthologs[seed][species]:
					codi = orthologs[seed][species]
					if codi in annotations[species]:
						print>>file,"[one-to-
one]\t"+seed+"\t"+orthologs[seed][species]+"\t"+annotations[species][codi]
				else:
					codis = orthologs[seed][species].split(",")
					go_terms = ""
					for codi in codis:
						if codi in annotations[species]:
							if go_terms == "":
								go_terms = annotations[species][codi]
							else:
								gos = annotations[species][codi].split(" || ")
								for go in gos:
									if go not in go_terms:
										go_terms = go_terms + " || " + go
					if go_terms != "":
						print >>file,"[one-to-
many]\t"+seed+"\t"+orthologs[seed][species]+"\t"+go_terms
		else:
			if species in orthologs[seed]:
				if "," not in orthologs[seed][species]:
					codi = orthologs[seed][species]
					if codi in annotations[species]:
						print >>file,"[many-to-
one]\t"+seed+"\t"+orthologs[seed][species]+"\t"+annotations[species][codi]
				else:
					codis = orthologs[seed][species].split(",")
					go_terms = ""
					for codi in codis:
						if codi in annotations[species]:
							if go_terms == "":
								go_terms = annotations[species][codi]
							else:
								gos = annotations[species][codi].split(" || ")
								for go in gos:
									if go not in go_terms:
										go_terms = go_terms + " || " + go
					if go_terms != "":
						print >>file,"[many-to-
many]\t"+seed+"\t"+orthologs[seed][species]+"\t"+go_terms
	file.close()
```

4. RESULTADOS

4.1. O filoma de *S. mansoni*

O filoma reconstruído neste trabalho é derivado da análise comparativa do conjunto de proteínas codificadas pelo genoma de *S. mansoni* com outros 16 organismos completamente sequenciados (Quadro 1).

Os conjuntos de proteínas com similaridade de sequência significativa (E-value $<10^{-3}$) e com cobertura mínima desejada (33%), totalizando 8.818 proteínas (66%) do proteoma predito de *S. mansoni*, foram alinhados, filtrados e deram origem a um total de 8.818 árvores filogenéticas. A ausência de árvore para as demais 4.467 proteínas (33,62%) justifica-se pelo fato destas não atenderem aos critérios de seleção descritos anteriormente (Materiais e Métodos, item 3.2). Das 8.818 árvores, 2.534 (29%) correspondem a árvores cujas proteínas em análise possuem função desconhecida, sendo anotadas como proteínas hipotéticas ou expressas e não possuem termos do *Gene Ontology* (GO) conforme descrito no *SchistoDB* (Zerlotini *et al.*, 2009). Na Figura 8 é possível observar a distribuição das proteínas hipotéticas ou expressas (violeta) de *S. mansoni* agrupadas de acordo com o tamanho das sequências proteicas em aminoácidos, bem como as respectivas árvores filogenéticas (verde) reconstruídas para estas proteínas.

Todos os alinhamentos e árvores filogenéticas resultantes desta análise foram depositadas no *PhylomeDB* (Huerta-Cepas *et al.*, 2008) e podem ser acessadas facilmente através do website http://phylomedb.org. Como as árvores e alinhamentos foram gerados automaticamente, em caso de análises mais detalhadas sobre famílias de proteínas específicas, é recomendado verificar a qualidade dos alinhamentos das sequências proteicas para julgar a qualidade dos dados. Se necessário, os alinhamentos podem ser expandidos ou refinados.

A mineração dos dados foi realizada via consultas em SQL no banco de dados relacional *SchistoPhylomeSQL* (Figura 9).

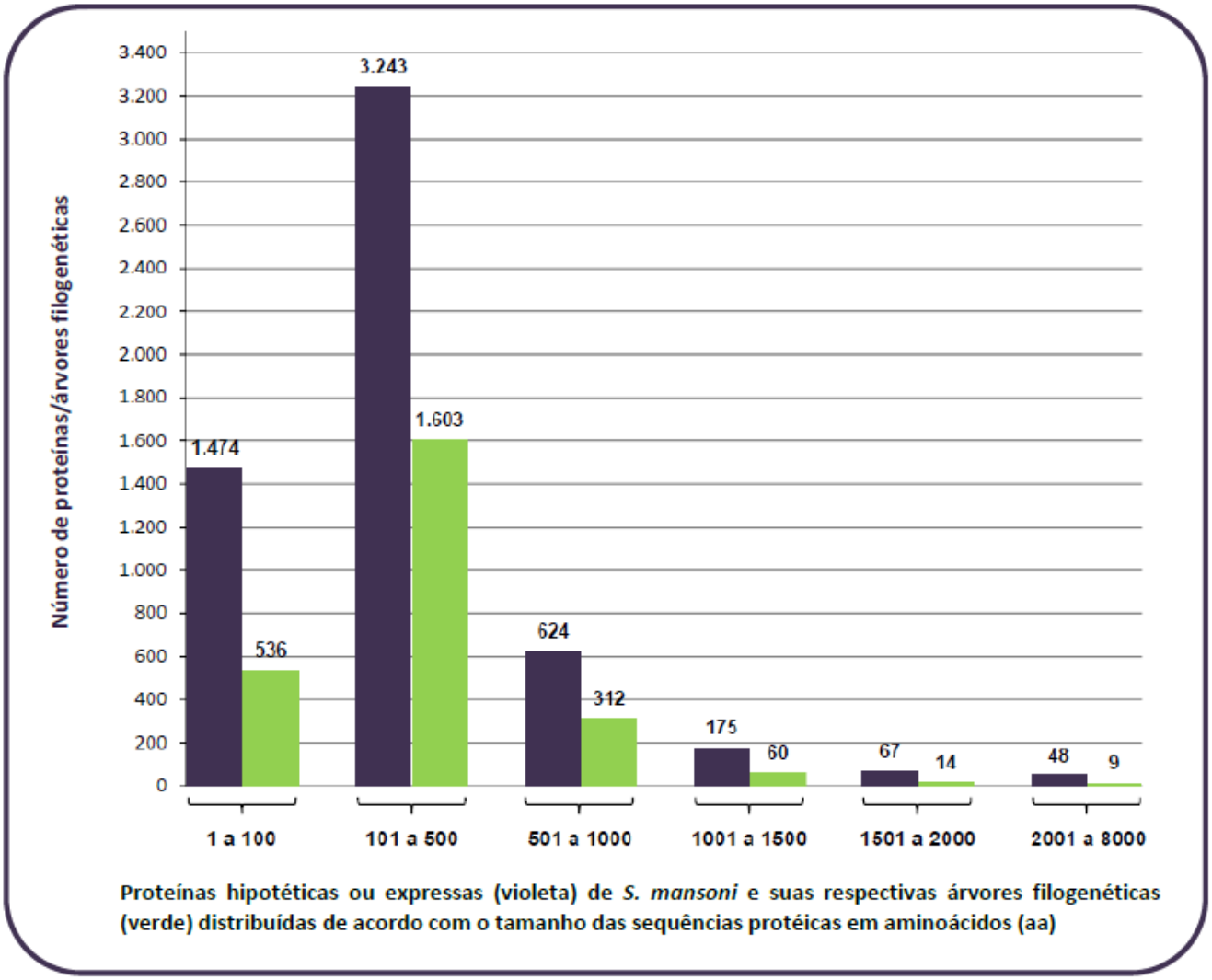


Figura 8: Distribuição das proteínas hipotéticas ou expressas (violeta) de *S. mansoni* e suas respectivas árvores filogenéticas (verde) no *PhylomeDB* de acordo com o tamanho das sequências proteicas em aminoácidos (aa).

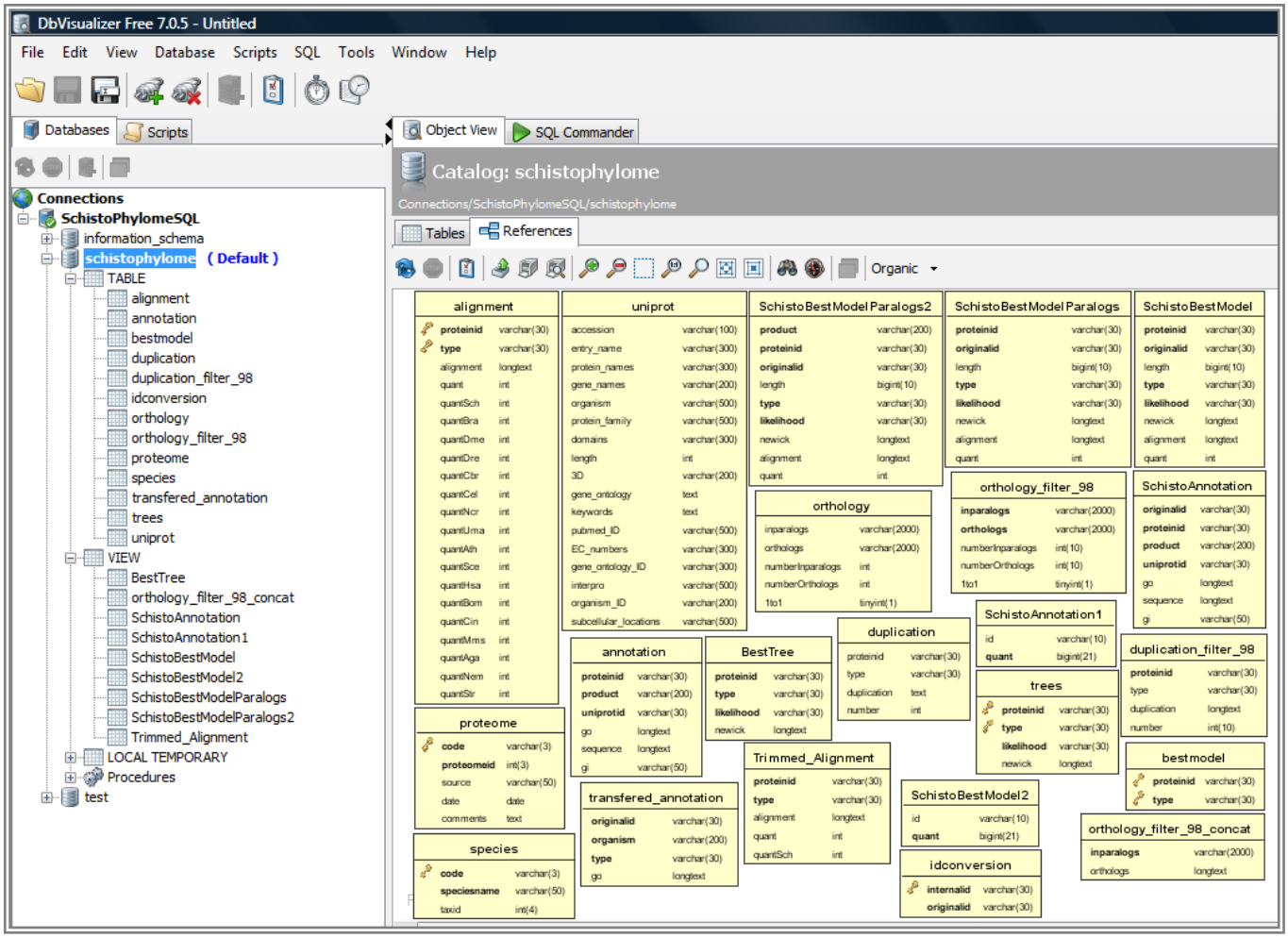


Figura 9: Interface gráfica do SchistoPhylomeSQL através do DbVisualizer.

4.2. Eliminação de produtos de *splicing* alternativo e sequências redundantes

A versão do proteoma predito de *S. mansoni* utilizada na reconstrução do filoma, disponível no banco de dados *SchistoDB* 2.0 (Zerlotini *et al.*, 2009), possui 1.282 produtos de *splicing* alternativo além de sequências redundantes (98% de identidade ou mais). Exemplos de produtos de *splicing* alternativo entre as sequências do proteoma predito de *S. mansoni* podem ser vistos no Quadro 5.

Para filtrar a redundância por similaridade, através do ETE (*Enviroment for tree Exploration*) (Huerta-Cepas *et al*., 2010a), conectamos no *PhylomeDB* e selecionamos as 6.118 árvores filogenéticas e seus respectivos alinhamentos que contêm, no mínimo, duas sequências de *S. mansoni.* A matriz de porcentagem de identidade de cada alinhamento foi obtida utilizando o programa trimAl (Capella-Gutierrez *et al.*, 2009). Um exemplo desta

matriz pode ser visualizado no Quadro 6. Em seguida um programa escrito em Python identificou as sequências de *S. mansoni* com porcentagem de identidade igual ou maior a 98% e selecionou apenas uma delas, descartando as demais. Exemplos de pares de sequências proteicas que apresentam pelo menos 98% de identidade estão ilustrados no Quadro 6. A escolha por eliminar sequências com altos valores de identidade ocorreu em função da presença de um grande número de identificadores com sequências nucleotídicas e proteicas altamente conservadas. Estas sequências estão localizadas em diferentes *scaffolds* e/ou posições no genoma, mas não existe nenhum tipo de informação adicional para os respectivos identificadores, os quais tampouco podem ser rastreados no mapa genético de *S. mansoni* (Criscione *et al.,* 2009). Na tentativa de identificar domínios protéicos nestas sequências, foram realizadas buscas no *Pfam* (Fiin *et al.*, 2010), um banco de dados de domínios protéicos (http://pfam.sanger.ac.uk). Entretanto, nenhum domínio foi identificado. No *UniProt* (The UniProt Consortium, 2010), um banco de dados de sequências proteicas (http://www.uniprot.org), todas estas sequências são agrupadas em um único número de acesso, sugerindo redundância no banco de dados. Exemplos destas sequências proteicas podem ser visualizados no Quadro 7.

Todas as sequências eliminadas foram devidamente documentadas. O conjunto de sequências eliminadas nesta etapa pode ser visualizado no Anexo I. No total, foram eliminadas 1.802 sequências (13,56 % do proteoma predito), sendo:

- ✓ 1003 sequências proteicas cuja função é predita ou conhecida experimentalmente;
- ✓ 498 sequências proteicas anotadas como *expressed protein;*
- ✓ 301 sequências proteicas anotadas como *hypothetical protein;*

} proteínas cuja função não é conhecida.

Quadro 5: Exemplos de proteínas codificadas pelo produto de *splicing* alternativo no PhylomeDB.

Identificador[1]	Identificador[2]	Comprimento (aa)[3]
Sch0001866	Smp_183020.17	75
Sch0001867	Smp_183020.16	75
Sch0001868	Smp_183020.13	75
Sch0001869	Smp_183020.21	75
Sch0001851	Smp_183020.20	80
Sch0001852	Smp_183020.3	80
Sch0001853	Smp_183020.10	80
Sch0001854	Smp_183020.8	80
Sch0001855	Smp_183020.18	80
Sch0001856	Smp_183020.5	80
Sch0001857	Smp_183020.9	80
Sch0001858	Smp_183020.4	80
Sch0001859	Smp_183020.12	80
Sch0001860	Smp_183020.14	80
Sch0001861	Smp_183020.19	80
Sch0001862	Smp_183020.11	80
Sch0001863	Smp_183020.2	80
Sch0001864	Smp_183020.7	80
Sch0001865	Smp_183020.1	80
Sch0001870	Smp_183020.15	80
Sch0001871	**Smp_183020.6**	**80**

1 – Identificador do produto de *splicing* alternativo da proteína Smp_183020 no PhylomeDB. 2 – Identificador da proteína Smp_183020 no SchistoDB. Os números após o ponto (.) indicam os diferentes transcritos. Todas as proteínas acima mencionadas estão anotadas no SchistoDB apenas como proteínas expressas (*expressed protein*). 3 – Comprimento da sequências proteicas em aminoácidos. Neste exemplo, a cor preta indica as sequências que foram eliminadas e a cor vermelha a sequência utilizada nas análises subsequentes.

Quadro 6: Matriz de porcentagem de identidade das sequências proteicas presentes no alinhamento múltiplo da proteína "semente" Sch0000871.

Matriz de porcentagem de identidade:							
Sch0005149	0	77,937	31,6176	14,3382	34	34	15,0735
Sch0005148	77,937	0	24,6418	11,1748	39,5	39,5	11,7479
Sch0003516	31,6176	24,6418	0	40,8602	12,2807	12,2807	43,0108
Sch0000871	14,3382	11,1748	40,8602	0	14,3216	14,3216	83,3333
Sch0006797	34	39.5	12,2807	14,3216	0	100	15,0754
Sch0006798	34	39.5	12,2807	14,3216	100	0	15,0754
Sch0002579	15,0735	11,7479	43,0108	83,3333	15,0754	15,0754	0
Porcentagem de identidade das sequências mais similares:							
Sch0005149	77,93	Sch0005148					
Sch0005148	77,93	Sch0005149					
Sch0003516	43.01	Sch0002579					
Sch0000871	83,33	Sch0002579					
Sch0006797	100,00	Sch0006798					
Sch0006798	100.00	**Sch0006797**					
Sch0002579	83,33	Sch0000871					

A sequência eliminada, de acordo com os parâmetros definidos, está destacada em verde. Para obter a matriz de porcentagem de identidade utilizamos o programa trimAl (Capella-Gutierrez *et al*., 2009).

Quadro 7: Exemplos de proteínas removidas do banco de dados em função do alto índice de identidade.

Identificador[1]	Localização[2]	Anotação Funcional[3]	UniProt[4]
Smp_188020	Smp_scaff011929: 760 - 1002	hypothetical protein	C1M1C7
Smp_114810	Smp_scaff011929: 1235 - 1477	hypothetical protein	C1M1C7
Smp_107780	Smp_scaff004236: 278 - 520	hypothetical protein	C1M1C7
Smp_107790	Smp_scaff004236: 753 - 995	hypothetical protein	C1M1C7
Smp_182720	Smp_scaff003783: 2644 - 2886	hypothetical protein	C1M1C7
Smp_107330	Smp_scaff003613: 474 - 716	hypothetical protein	C1M1C7
Smp_107650	Smp_scaff004053: 723 - 965	hypothetical protein	C1M1C7
Smp_184720	Smp_scaff007292: 160 - 402	hypothetical protein	C1M1C7
Smp_184730	Smp_scaff007292: 635 - 877	hypothetical protein	C1M1C7
Smp_112250	Smp_scaff008824: 89 - 331	hypothetical protein	C1M1C7
Smp_185490	Smp_scaff008690: 1222 - 1464	hypothetical protein	C1M1C7
Smp_111950	Smp_scaff008508: 187 - 429	hypothetical protein	C1M1C7
Smp_111960	Smp_scaff008508: 662 - 904	hypothetical protein	C1M1C7
Smp_185520	Smp_scaff008752: 842 - 1084	hypothetical protein	C1M1C7
Smp_185530	Smp_scaff008752: 1317 - 1559	hypothetical protein	C1M1C7
Smp_185540	Smp_scaff008752: 1792 - 2034	hypothetical protein	C1M1C7
Smp_111090	Smp_scaff007706: 251 - 493	hypothetical protein	C1M1C7
Smp_111100	Smp_scaff007706: 726 - 968	hypothetical protein	C1M1C7
Smp_110790	Smp_scaff007390: 594 - 836	hypothetical protein	C1M1C7
Smp_110970	Smp_scaff007624: 758 - 1000	hypothetical protein	C1M1C7
Smp_184950	Smp_scaff007624: 1708 - 1950	hypothetical protein	C1M1C7
Smp_110990	Smp_scaff007624: 2183 - 2425	hypothetical protein	C1M1C7
Smp_108070	Smp_scaff004529: 894 - 1136	hypothetical protein	C1M1C7
Smp_108090	Smp_scaff004529: 1844 - 2086	hypothetical protein	C1M1C7
Smp_184620	Smp_scaff007120: 772 - 1014	hypothetical protein	C1M1C7
Smp_184630	Smp_scaff007120: 1247 - 1489	hypothetical protein	C1M1C7
Smp_108520	Smp_scaff005069: 800 - 1042	hypothetical protein	C1M1C7
Smp_108730	Smp_scaff005443: 586 - 828	hypothetical protein	C1M1C7
Smp_108440	Smp_scaff004974: 1020 - 1262	hypothetical protein	C1M1C7
Smp_189060	Smp_scaff013234: 479 - 721	hypothetical protein	C1M1C7
Smp_117230	Smp_scaff015434: 76 - 318	hypothetical protein	C1M1C7
Smp_190490	Smp_scaff015434: 551 - 793	hypothetical protein	C1M1C7
Smp_190500	Smp_scaff015434: 1026 - 1268	hypothetical protein	C1M1C7
Smp_117270	Smp_scaff015482: 197 - 439	hypothetical protein	C1M1C7
Smp_117280	Smp_scaff015482: 672 - 914	hypothetical protein	C1M1C7
Smp_117500	Smp_scaff015754: 914 - 1156	hypothetical protein	C1M1C7
Smp_117520	Smp_scaff015754: 1864 - 2106	hypothetical protein	C1M1C7
Smp_117540	Smp_scaff015754: 2814 - 3056	hypothetical protein	C1M1C7

1 – Identificador da proteína no SchistoDB. 2 – Localização no genoma de *S. mansoni*. 3 – Anotação funcional disponível no SchistoDB. 4 – Código de acesso da sequência proteica no UniProt.

4.3. Análise filogenética das proteínas mitocondriais

Através do *pipeline* utilizado (Figura 5) foram obtidas árvores filogenéticas para 9 proteínas mitocondriais (COXI, COXII, COXIII, NAD1, NAD2, NAD3, NAD4, NAD5, NAD6, NAD4L, ATP6 e CYTB). As proteínas citocromo *b* (CYTB) (Figura 11) e citocromo oxidase subunidade I (COXI) (Figura 10) foram utilizadas como grupo controle para a reconstrução da história evolutiva de *S. mansoni*, em função da presença de homólogos destas proteínas em todas as espécies selecionadas para este estudo. CYTB é um componente da proteína integral de membrana do complexo bc1 que catalisa a transferência de elétrons da ubiquinona para o citocromo *c* na cadeia de transporte de elétrons mitocondrial (Gering *et al*., 2009) e COXI é o catalisador terminal da cadeia respiratória mitocondrial, é a maior subunidade das três citocromo oxidades e é uma das maiores proteínas codificadas pelo genoma mitocondrial (Kranthi *et al*., 2006).

Utilizando o programa FigTree (Rambaut, 2010), analisamos as árvores em formato gráfico e identificamos uma sequência homóloga à proteína de *S. mansoni* em cada espécie dos demais metazoários, exceto em *S. cerevisiae* onde estão presentes múltiplas proteínas homólogas, provavelmente originadas por duplicação gênica. A correta identificação de sequências homólogas das proteínas CYTB e COXI demonstra a capacidade de detecção acurada de relações de homologia pelo *pipeline* utilizado (Figura 5), validando a metodologia e aumentando a possibilidade de que a história evolutiva dos organismos reconstruída neste trabalho represente a realidade.

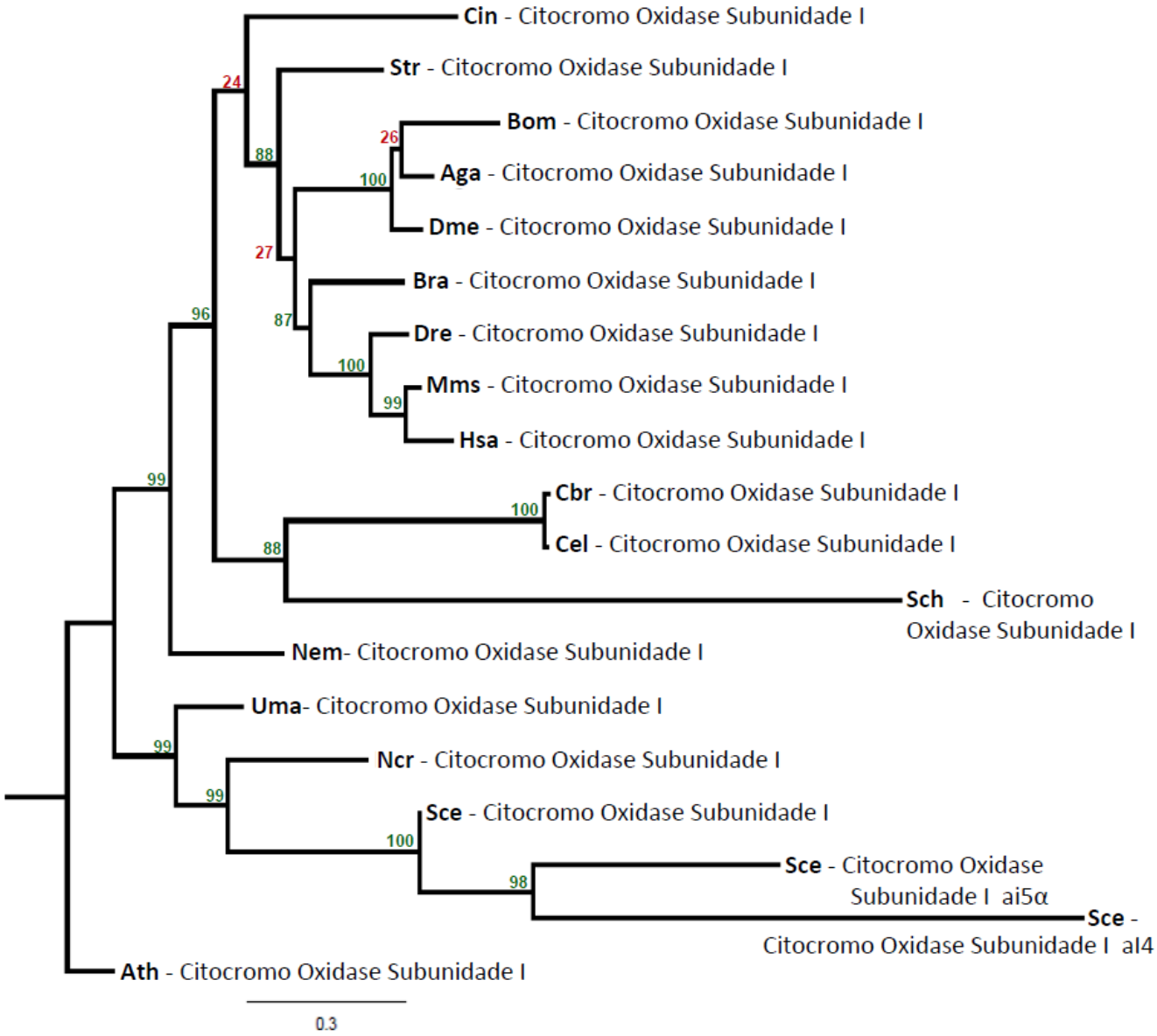


Figura 10: Árvore filogenética da subunidade I da proteína citocromo oxidase (COX I) reconstruída a partir do alinhamento de 510 aminoácidos pelo método de máxima verossimilhança. Segundo o critério de Akaike (Akaike, 1973), o modelo evolutivo que melhor se adequou aos dados é o BLOSUM62 cujo valor de loglk é igual a -11235.78170. Os valores de apoio dos agrupamentos apresentados na árvore foram obtidos através do *approximate Likelihood Ratio Test* (aLRT) (Guindon & Gascuel, 2003; Anisimova & Gascuel, 2006). Altos valores de apoio (acima de 80%) estão destacados em verde e os baixos em vermelho.

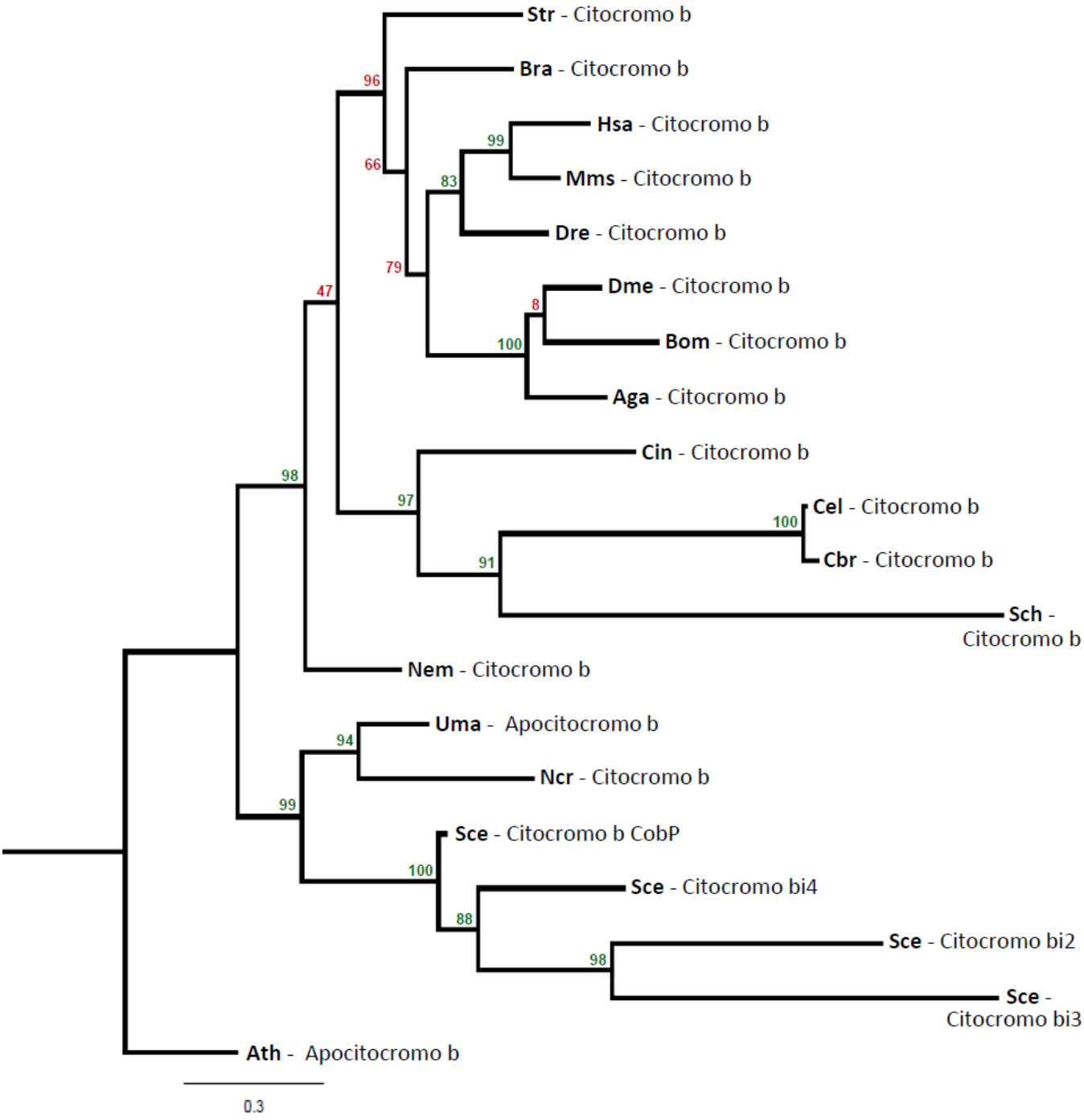


Figura 11: Árvore filogenética da proteína citocromo *b* (CYTB) reconstruída a partir do alinhamento de 364 aminoácidos pelo método de máxima verossimilhança. Segundo o critério de Akaike (Akaike, 1973), o modelo evolutivo que melhor se adequou aos dados é o BLOSUM62 cujo valor de loglk é igual a -10898.73909. Os valores de apoio dos agrupamentos apresentados na árvore foram obtidos através do *approximate Likelihood Ratio Test* (aLRT) (Guindon & Gascuel, 2003; Anisimova & Gascuel, 2006). Altos valores de apoio (acima de 80%) estão destacados em verde e os baixos em vermelho.

4.4. Inferência da função gênica por métodos filogenômicos

A predição funcional de 946 proteínas de *S. mansoni* foi realizada utilizando os dados das relações evolutivas (árvores filogenéticas) previamente obtidos através do algoritmo que realiza a predição de relações de ortologia e paralogia (Materiais e Métodos, item 3.7). Os resultados desta análise foram armazenados em arquivos texto (e.g. Figura 7) e exemplos de ortólogos preditos e relações de homologia podem ser obtidos nos Quadros 8 e 9, respectivamente.

Quadro 8: Exemplo de ortólogos preditos para a sequência proteica de *S. mansoni* Sch0005783.

Identificador[1]	Ortólogos[2]
Sch0005783	Cbr0012032
Sch0005783	Nem0015259
Sch0005783	Dme0001733
Sch0005783	Cel0010302
Sch0005783	Cin0014271
Sch0005783	Mms0010746
Sch0005783	Aga0008515
Sch0005783	Hsa0011772
Sch0005783	Ath0021423,Ath0008631
Sch0005783	Str0017449,Str0019145,Str0007156,Str0004664
Sch0005783	Dre0015430
Sch0005783	Bra0045297, Bra0029501
Sch0005783	Bom0005360
Sch0005783	Sch0005783

1 – Identificador da proteína de *S. mansoni* utilizada como “semente” na construção da árvore filogenética. 2 – Ortólogos preditos.

Quadro 9: Exemplos de relações de homologia entre sequências proteicas de *S. mansoni* e *Homo sapiens.*

Relação de Homologia[1]	Identificadores[2]	Identificadores[3]
um-pra-um	Sch0013240	Hsa0022755
um-pra-muitos	Sch0013084	Hsa0007244,Hsa0017989
muitos-pra-um	Sch0011633, Sch0008557, Sch0004019, Sch0011098, Sch0004424, Sch0011634, Sch0011949, Sch0004850	Hsa0021925
muitos-pra-muitos	Sch0013219, Sch0013220	Hsa0012407, Hsa0012409, Hsa0012408, Hsa0012410

1 – Relações de homologia entre sequências proteicas de *S. mansoni* e *Homo* sapiens. 2 – Identificadores das proteínas de *S. mansoni* no PhylomeDB. 3 – Identificadores das proteínas de *H. sapiens* no PhylomeDB.

Estes subconjuntos de dados são derivados de análises realizadas separadamente para cada espécie, efetuando a transferência de anotação funcional para as proteínas de *S. mansoni* com base nas relações evolutivas, a partir de potenciais ortólogos com função conhecida na outra espécie.

Em síntese, esta análise gerou um total de 53.189 transferências de anotação funcional para 5.507 sequências de proteínas de *S. mansoni*. Exemplos desta transferência podem ser visualizados no Quadro 10. Do total de 5.631 proteínas cuja função não havia sido predita ou confirmada experimentalmente, foi possível transferir anotação funcional para 946 (7% do genoma predito). As predições funcionais transferidas para estas sequências podem ser visualizadas no Anexo II. No caso de proteínas para as quais dados de predição funcional já haviam sido obtidos anteriormente, estes resultados serão úteis para confirmar, complementar ou até mesmo alterar a predição funcional destas sequências proteicas (Quadro 11). A confiabilidade da transferência é especialmente alta para 677 (destacadas em vermelho no Anexo II) das 946 sequências, em função da predição ter sido baseada no subconjunto de ortologias um-pra-um, significando que há apenas um ortólogo destas sequências na espécie fonte da anotação. Um exemplo de transferência um-pra-um de *Caenorhabidtis briggsae* para *S. mansoni* pode ser visualizado na Figura 12 que mostra a reconstrução filogenética da proteína Sch0002900 e seus potenciais homólogos nas demais espécies analisadas neste trabalho (Quadro 1). A análise desta árvore resultou em 13

transferências de anotação funcional de ortólogos com função conhecida para a proteína "semente" de *S. mansoni.* Neste exemplo, para ilustrar um caso de transferência um-pra-um escolhemos a anotação funcional transferida de *Caenorhabidtis briggsae* apenas por se tratar do organismo mais próximo taxonomicamente cujo ortólogo possui função conhecida. No lado direito da árvore, em frente ao código do gene é possível identificar os termos *Gene Ontology* (GO) transferidos de *C. briggsae* para *S. mansoni*, bem como é possível identificar os termos GO para outros ortólogos, dando maior apoio à transferência.

Dentre as árvores filogenéticas resultantes, oito possuem apenas sequências de *S. mansoni* e *H. sapiens* (Smp_157140.2, Smp_037460.1, Smp_133530.2, Smp_125460, Smp_192700, Smp_102210, Smp_074250, Smp_165020.2) e em outros casos, embora existam sequências de outros organismos na árvore, os identificadores do parasito e do hospedeiro definitivo estão presentes de forma isolada no mesmo clado (Smp_171690, Smp_142810, Smp_078560, Smp_159420, Smp_159920, Smp_130240). Um exemplo interessante pode ser visualizado na árvore Smp_159920, onde uma caderina (molécula de adesão celular) que possui vários parálogos em *S. mansoni* esta posicionada em um clado com uma única caderina de *H. sapiens* (ENSP00000262150) (Figura 13).

Quadro 10: Exemplos de transferência de anotação funcional de uma proteína ortóloga com função conhecida para a proteína hipotética ou expressa de *S. mansoni*.

Identificador[1]	Código[2]	Relação de Homologia[3]	Termos GO[4]
Smp_015490	Hsa	[muitos-pra-um]	GO:0006508 proteolysis || GO:0008237 metallopeptidase activity || GO:0008270 zinc ion binding || GO:0008450 O-sialoglycoprotein endopeptidase activity || GO:0046872 metal ion binding
Smp_015490	Dme	[muitos-pra-um]	GO:0006508 proteolysis || GO:0008270 zinc ion binding || GO:0008450 O-sialoglycoprotein endopeptidase activity
Smp_015490	Cel	[muitos-pra-um]	GO:0004175 endopeptidase activity
Smp_015490	Aga	[muitos-pra-um]	GO:0006508 proteolysis || GO:0008270 zinc ion binding || GO:0008450 O-sialoglycoprotein endopeptidase activity
Smp_015490	Nem	[muitos-pra-um]	GO:0006508 proteolysis || GO:0008450 O-sialoglycoprotein endopeptidase activity || GO:0008270 zinc ion binding
Smp_015490	Mms	[muitos-pra-um]	GO:0006508 proteolysis || GO:0008237 metallopeptidase activity || GO:0008270 zinc ion binding || GO:0008450 O-sialoglycoprotein endopeptidase activity || GO:0046872 metal ion binding
Smp_015490	Str	[muitos-pra-muitos]	GO:0006508 proteolysis || GO:0004222 metalloendopeptidase activity || GO:0008270 zinc ion binding || GO:0004175 endopeptidase activity
Smp_015490	Dre	[muitos-pra-um]	GO:0006508 proteolysis || GO:0008270 zinc ion binding || GO:0008450 O-sialoglycoprotein endopeptidase activity
Smp_027740	Hsa	[muitos-pra-um]	GO:0000166 nucleotide binding || GO:0004798 thymidylate kinase activity || GO:0005524 ATP binding || GO:0006233 dTDP biosynthetic process || GO:0006235 dTTP biosynthetic process || GO:0006259 DNA metabolic process || GO:0007049 cell cycle || GO:0008283 cell proliferation || GO:0009165 nucleotide biosynthetic process || GO:0016301 kinase activity || GO:0016740 transferase activity
Smp_027740	Dme	[muitos-pra-um]	GO:0004798 thymidylate kinase activity || GO:0005515 protein binding || GO:0005524 ATP binding || GO:0006233 dTDP biosynthetic process || GO:0006235 dTTP biosynthetic process
Smp_027740	Sce	[muitos-pra-um]	GO:0000166 nucleotide binding || GO:0004798 thymidylate kinase activity || GO:0005524 ATP binding || GO:0005634 nucleus || GO:0005737 cytoplasm || GO:0006227 dUDP biosynthetic process || GO:0006233 dTDP biosynthetic process || GO:0006235 dTTP biosynthetic process || GO:0006261 DNA-dependent DNA replication || GO:0006276 plasmid maintenance || GO:0006280 mutagenesis || GO:0006281 DNA repair || GO:0009041 uridylate kinase activity || GO:0009165 nucleotide biosynthetic process || GO:0016301 kinase activity || GO:0016740 transferase activity || GO:0042802 identical protein binding
Smp_027740	Ncr	[muitos-pra-um]	GO:0004798 thymidylate kinase activity || GO:0005524 ATP binding || GO:0006233 dTDP biosynthetic process || GO:0016301 kinase activity
Smp_027740	Cel	[muitos-pra-um]	GO:0002119 nematode larval development || GO:0009792 embryonic development ending in birth or egg hatching || GO:0040007 growth
Smp_027740	Mms	[muitos-pra-um]	GO:0004798 thymidylate kinase activity || GO:0005524 ATP binding || GO:0006233 dTDP biosynthetic process || GO:0006235 dTTP biosynthetic process || GO:0006259 DNA metabolic process || GO:0007049 cell cycle || GO:0008283 cell proliferation || GO:0016301 kinase activity
Smp_027740	Nem	[muitos-pra-um]	GO:0004798 thymidylate kinase activity || GO:0005524 ATP binding || GO:0006233 dTDP biosynthetic process || GO:0006235 dTTP biosynthetic process
Smp_027740	Uma	[muitos-pra-um]	GO:0004798 thymidylate kinase activity || GO:0005524 ATP binding || GO:0006233 dTDP biosynthetic process
Smp_027740	Aga	[muitos-pra-um]	GO:0004798 thymidylate kinase activity || GO:0005515 protein binding || GO:0005524 ATP binding || GO:0006233 dTDP biosynthetic process || GO:0006235 dTTP biosynthetic process
Smp_027740	Bra	[muitos-pra-muitos]	GO:0004798 thymidylate kinase activity ||GO:0005524 ATP binding ||GO:0006233 dTDP biosynthesis ||GO:0006235 dTTP biosynthesis || GO:0006235 dTTP biosynthesis ||GO:0006233 dTDP biosynthesis ||GO:0005524 ATP binding ||GO:0004798 thymidylate kinase activity

1 – Identificador da proteína hipotética ou expressa de *S. mansoni.* 2 – Código da espécie fonte da anotação funcional. 3 – Relação de homologia entre a(s) sequência(s) proteicas de *S. mansoni* e a(s) sequencia(s) proteicas da outra espécie. 4 – Termos GO (*Gene Ontology*) transferidos para a sequência proteica de *S. mansoni*.

Quadro 11: Exemplos de avanço na anotação funcional do proteoma de *S. mansoni* utilizando a abordagem filogenômica.

Identificador[1]	Identificador[2]	GO[3]	GO[4]
Smp_073080.1	Sch0005651	GO:0003676 - F nucleic acid binding	GO:0003723 RNA binding
Smp_161920	Sch0005609	GO:0005515 - F protein binding	GO:0000166 nucleotide binding || GO:0005198 structural molecule activity || GO:0005200 structural constituent of cytoskeleton || GO:0005515 protein binding || GO:0005524 ATP binding || GO:0005856 cytoskeleton || GO:0005884 actin filament
Smp_083650	Sch0004506	GO:0006520 - P amino acid metabolic process	GO:0003824 catalytic activity || GO:0004123 cystathionine gamma-lyase activity || GO:0006520 cellular amino acid metabolic process || GO:0016829 lyase activity || GO:0030170 pyridoxal phosphate binding
Smp_084380	Sch0004433	GO:0016787 - F hydrolase activity	GO:0004518 nuclease activity || GO:0004519 endonuclease activity || GO:0005634 nucleus || GO:0006397 mRNA processing || GO:0008270 zinc ion binding || GO:0016787 hydrolase activity || GO:0046872 metal ion binding
Smp_168770	Sch0004271	GO:0031072 - F heat shock protein binding	GO:0005743 mitochondrial inner membrane || GO:0065002 intracellular protein transmembrane transport
Smp_175560	Sch0003105	GO:0004194 - F pepsin A activity || GO:0006508 - P proteolysis	GO:0004190 aspartic-type endopeptidase activity || GO:0004194 pepsin A activity || GO:0006508 proteolysis || GO:0008233 peptidase activity || GO:0009049 aspartic-type signal peptidase activity || GO:0016787 hydrolase activity
Smp_094530	Sch0003386	GO:0005737 - C cytoplasm || GO:0019478 - P D-amino acid catabolic process	GO:0005737 cytoplasm || GO:0016788 hydrolase activity, acting on ester bonds || GO:0019478 D-amino acid catabolic process
Smp_194230	Sch0000100	GO:0005975 - P carbohydrate metabolic process	GO:0004553 hydrolase activity, hydrolyzing O-glycosyl compounds || GO:0005975 carbohydrate metabolic process
Smp_120660	Sch0000139	GO:0005975 - P carbohydrate metabolic process	GO:0005975 carbohydrate metabolism ||GO:0004553 hydrolase activity, hydrolyzing O-glycosyl compounds
Smp_120320	Sch0000185	GO:0003723 - F RNA binding || GO:0006396 - P RNA processing	GO:0000381 regulation of alternative nuclear mRNA splicing, via spliceosome || GO:0003723 RNA binding || GO:0005515 protein binding || GO:0005634 nucleus || GO:0006396 RNA processing || GO:0006464 protein modification process
Smp_120080.2	Sch0000230	GO:0006807 - P nitrogen compound metabolic process	GO:0016810 hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds || GO:0006807 nitrogen compound metabolic process

1 – Identificador da proteína de *S. mansoni* no SchistoDB. 2 – Identificador da proteína de *S. mansoni* no PhylomeDB. 3 – Termos GO (*Gene Ontology*) disponíveis previamente para a proteína no SchistoDB. 4 – Termos GO (*Gene Ontology*) transferidos para a sequência proteica através da análise filogenômica.

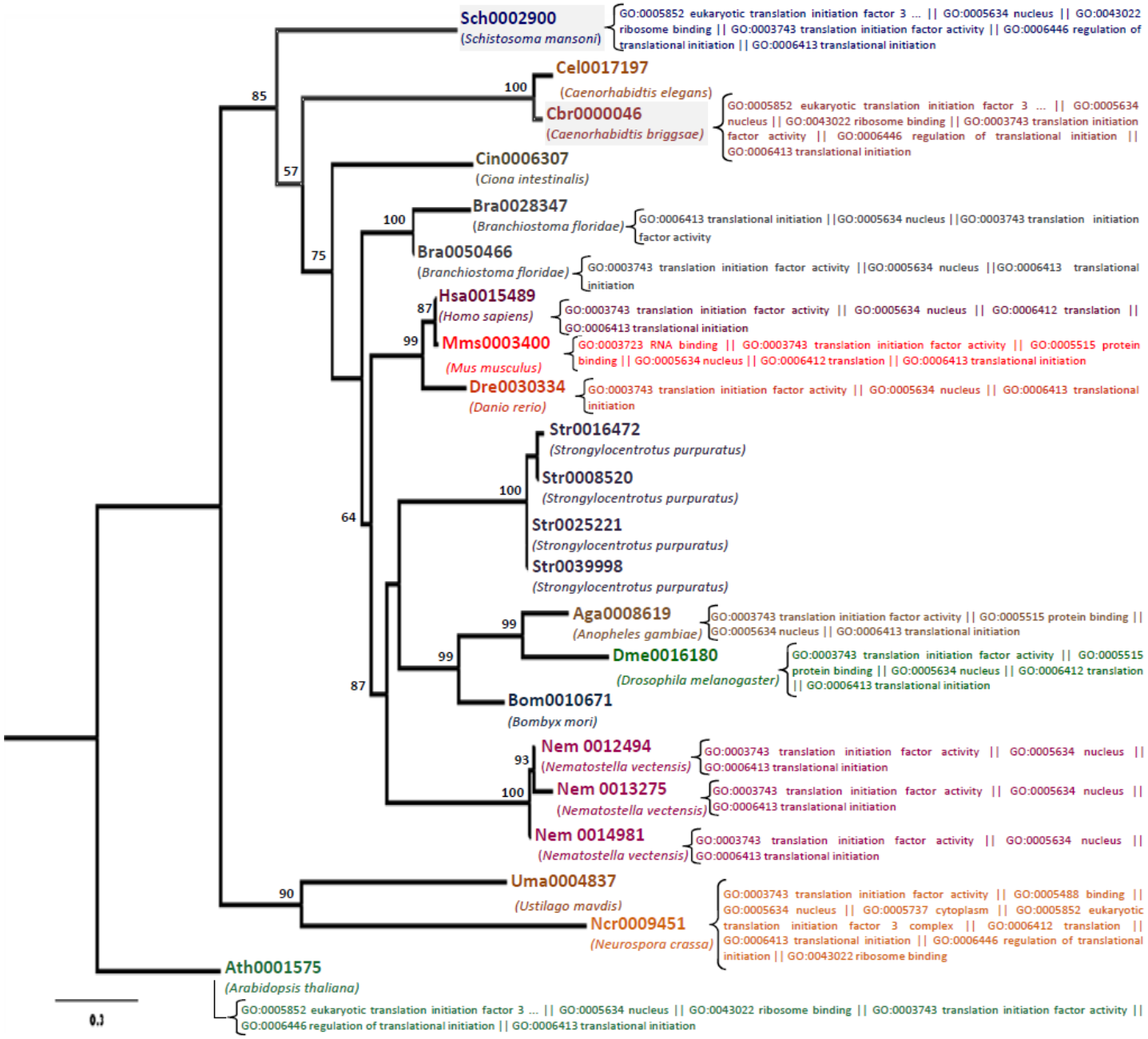


Figura 12: Exemplo de transferência de anotação funcional um-pra-um de *Caenorhabidtis briggsae* para a proteína 'semente' Sch0002900 de *Schistosoma mansoni* (ambos destacados em cinza). As sequências proteicas estão representadas pelo identificador interno no *PhylomeDB*. Os valores de apoio dos agrupamentos e os termos *Gene Ontology* (GO) de cada proteína também estão indicados.

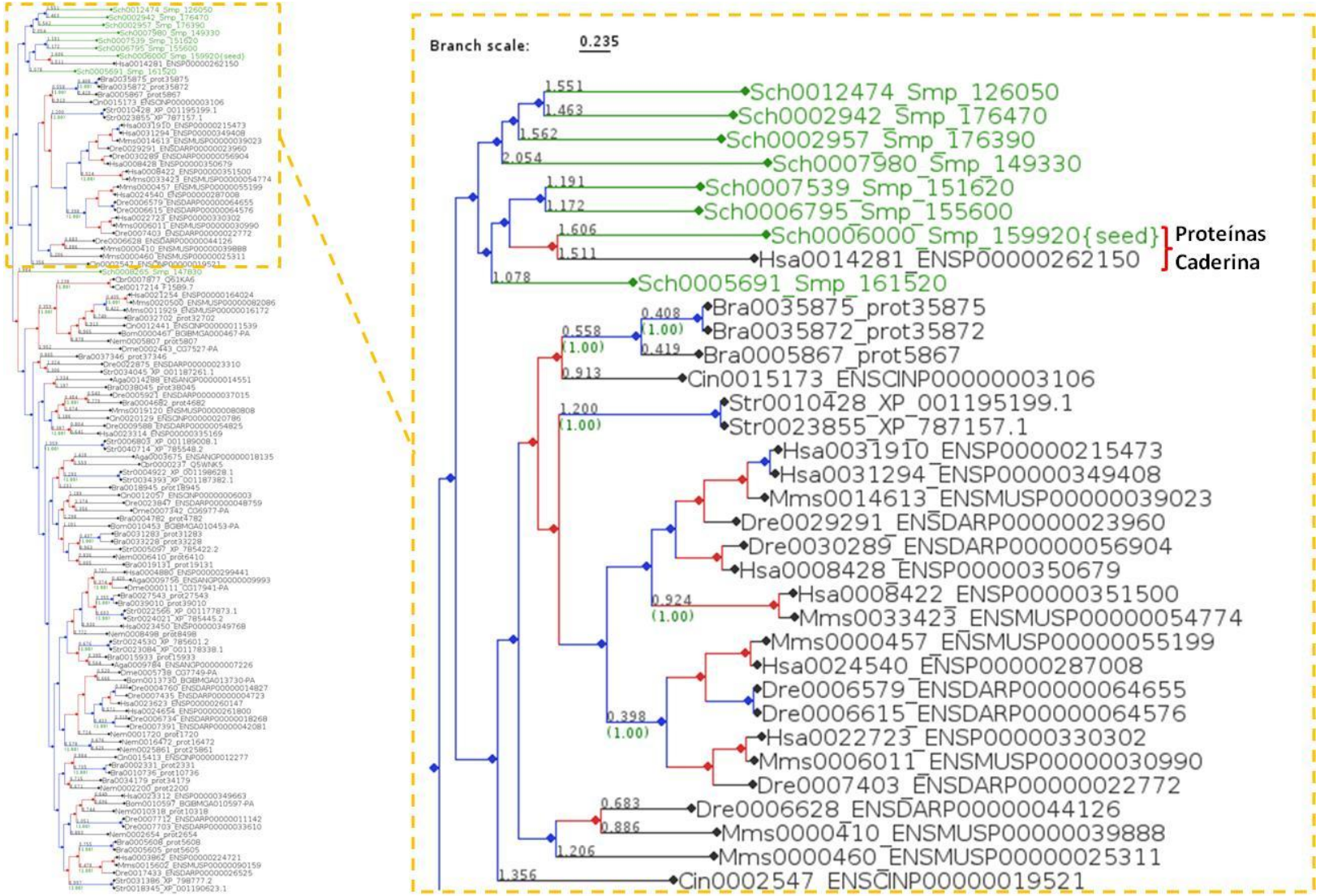


Figura 13: Árvore filogenética reconstruída para a proteína Smp_159920. A proteína "semente" é uma caderina (molécula de adesão celular) que possui vários parálogos em *S. mansoni*. É possível observar que esta proteína está posicionada em um clado com uma única caderina de *H. sapiens*, embora existam nesta árvore sequências de organismos mais próximos taxonomicamente tanto de *S. mansoni* quanto de *H. sapiens.*

4.5 – Duplicações gênicas em *S. mansoni*

Utilizando o algoritmo que realiza a predição de parálogos (Huerta-Cepas *et al*., 2007), através dos dados do filoma de *S. mansoni*, foi possível detectar as duplicações gênicas que ocorreram no genoma deste parasito. 1.805 genes (13,59% do proteoma predito) possuem, no mínimo, 1 parálogo no genoma (Anexo III). A maioria das expansões (1.716) são pequenas, resultando em um total de 2 a 10 *in-paralogs*, parálogos que surgiram após o processo de especiação (Remm *et al.*, 2001). As demais expansões são mais significativas resultando em grupos maiores, de 11 a 50 membros (61 genes) ou de 51 a 89 membros (28 genes).

O algoritmo que realiza a predição de relações de paralogia considera como duplicação apenas as unidades taxonômicas que estão conectadas a um nó marcado como evento de duplicação. A análise tem início na proteína "semente" e término na raiz da árvore filogenética. Por este motivo, as duplicações aqui mencionadas correspondem, em sua grande maioria, apenas a eventos de duplicação que ocorreram após a divergência entre *S. mansoni* e os demais táxons selecionados (Quadro 1). Um exemplo de como o algoritmo atua pode ser obtido analisando a árvore filogenética da proteína "semente" Sch0013135 (Figura 14). O identificador Sch0009447 não é contabilizado como parálogo em função deste estar localizado na porção inferior à raiz da árvore.

Do total de 1.805 identificadores (14% do proteoma predito) para os quais existe pelo menos um parálogo no genoma de *S. mansoni*, 477 (3,6% do proteoma predito) são anotados como genes cuja função é desconhecida. Destes últimos, 355 não possuem termos *Gene Ontology*. Neste trabalho foi possível predizer a função de 212 destes genes (Anexo III), inclusive de codificantes de enzimas com atividade kinase (Smp_027740) e fucosiltransferase (Smp_138740). Nas Figuras 15 a 17 é possível visualizar o agrupamento dos termos GO mais frequentes na anotação funcional das proteínas duplicadas para as quais a função foi predita neste trabalho.

Na categoria "processo biológico", prevalecem processos como proteólise, metabolismo, transporte de proteínas, adesão celular e reparo de DNA (Figura 15). Na categoria "componente celular", por sua vez, são mais frequentes os produtos gênicos localizados na membrana, núcleo, integrais à membrana, citoplasma, complexo de Golgi, mitocôndria, espaço extracelular e espaço intracelular (Figura 16). Por fim, na categoria "função molecular" destacam-se os produtos que se ligam ao DNA e ATP, atividade catalítica, ligação a íons zinco, ligação e atividade hidrolase (Figura 17).

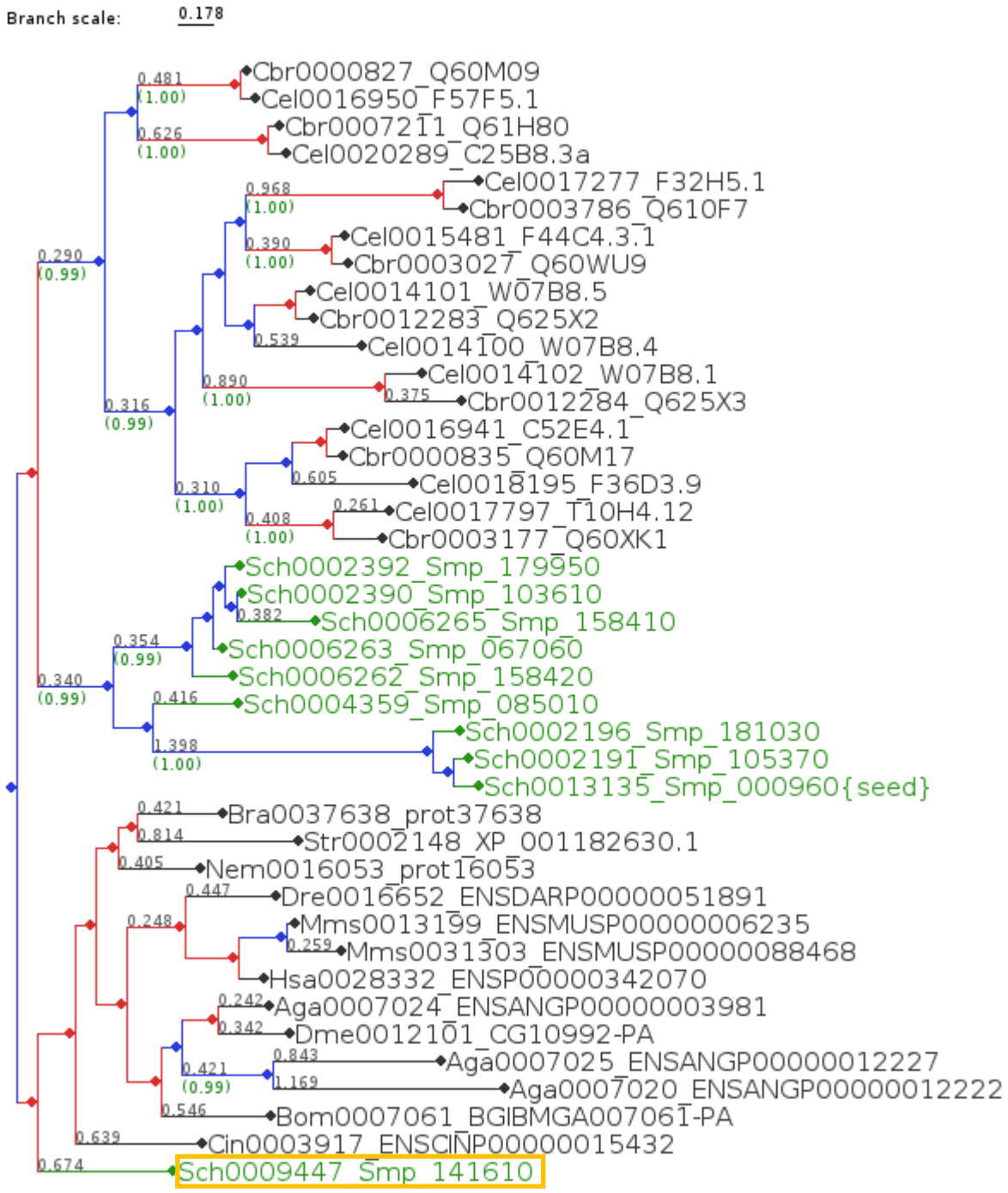


Figura 14: Árvore filogenética da proteína "semente" Sch0013135 e seus homólogos em 12 organismos. As proteínas de *S. mansoni* estão destacadas em verde. Neste exemplo, a proteína Sch0009447_Smp_141610 (destaque em amarelo) não é contabilizada como paráloga das outras proteínas de *S. mansoni* presentes nesta árvore em função desta estar localizada na porção inferior à raiz da árvore.

Dentre os genes duplicados em *S. mansoni* com função descrita previamente destacam-se aqueles amplamente estudados como alvos para drogas, vacina e diagnóstico contra a esquistossomose, como tetraspaninas, fucosiltransferases, leishmanolisinas e Scp-like (Marques *et al.*, 2001; Tran *et al.*, 2006; McManus *et al.*, 2008; DeMarco & Verjovski-Almeida, 2009; Berriman *et al.,* 2009; Fitzpatrick *et al.*, 2009).

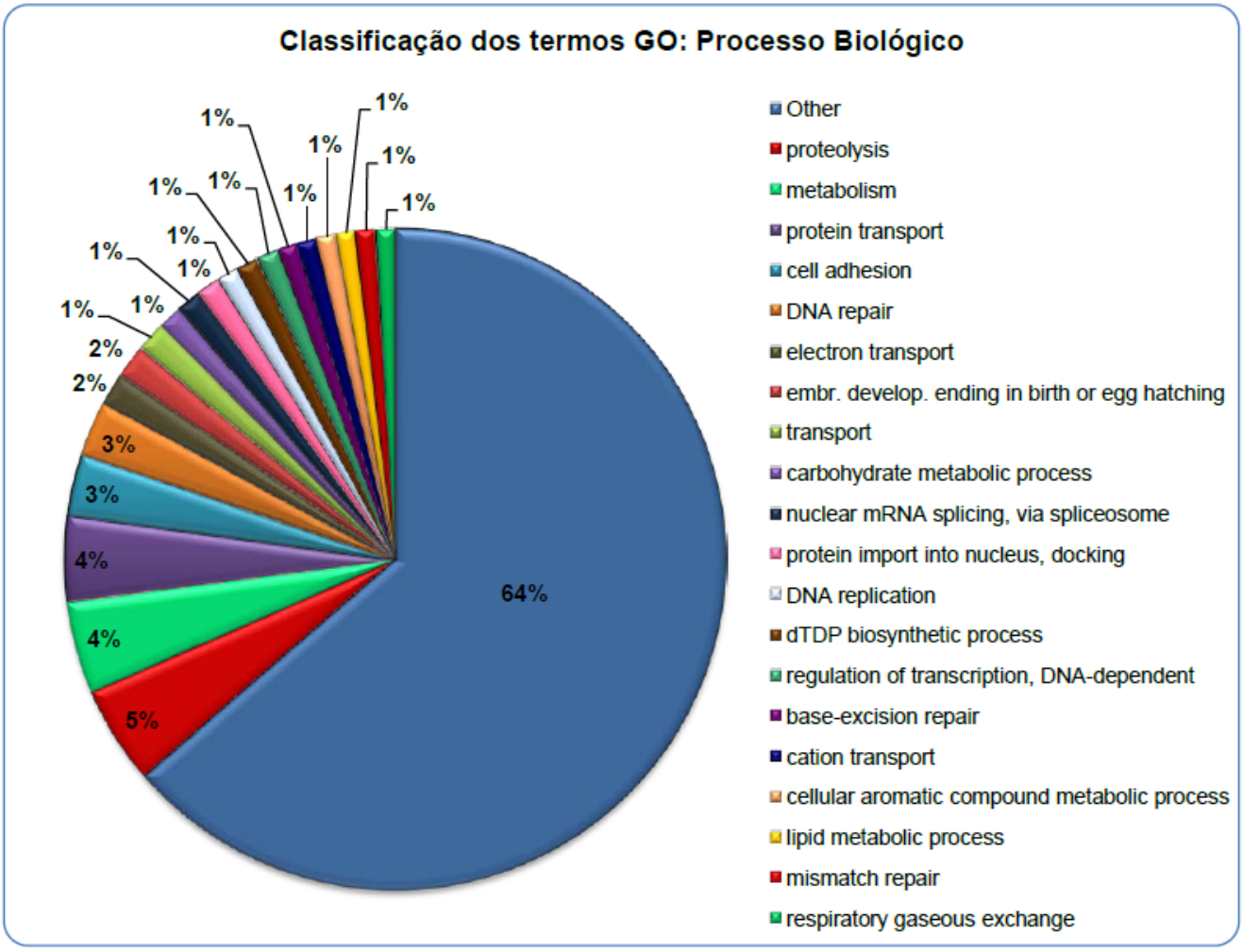


Figura 15: Agrupamento dos termos GO mais frequentes, da categoria "processos biológicos", transferidos para as proteínas duplicadas para as quais a função foi predita. A ordem dos termos no gráfico obedece a disposição visualizada na legenda, tendo início em "Other" e seguindo no sentido horário.

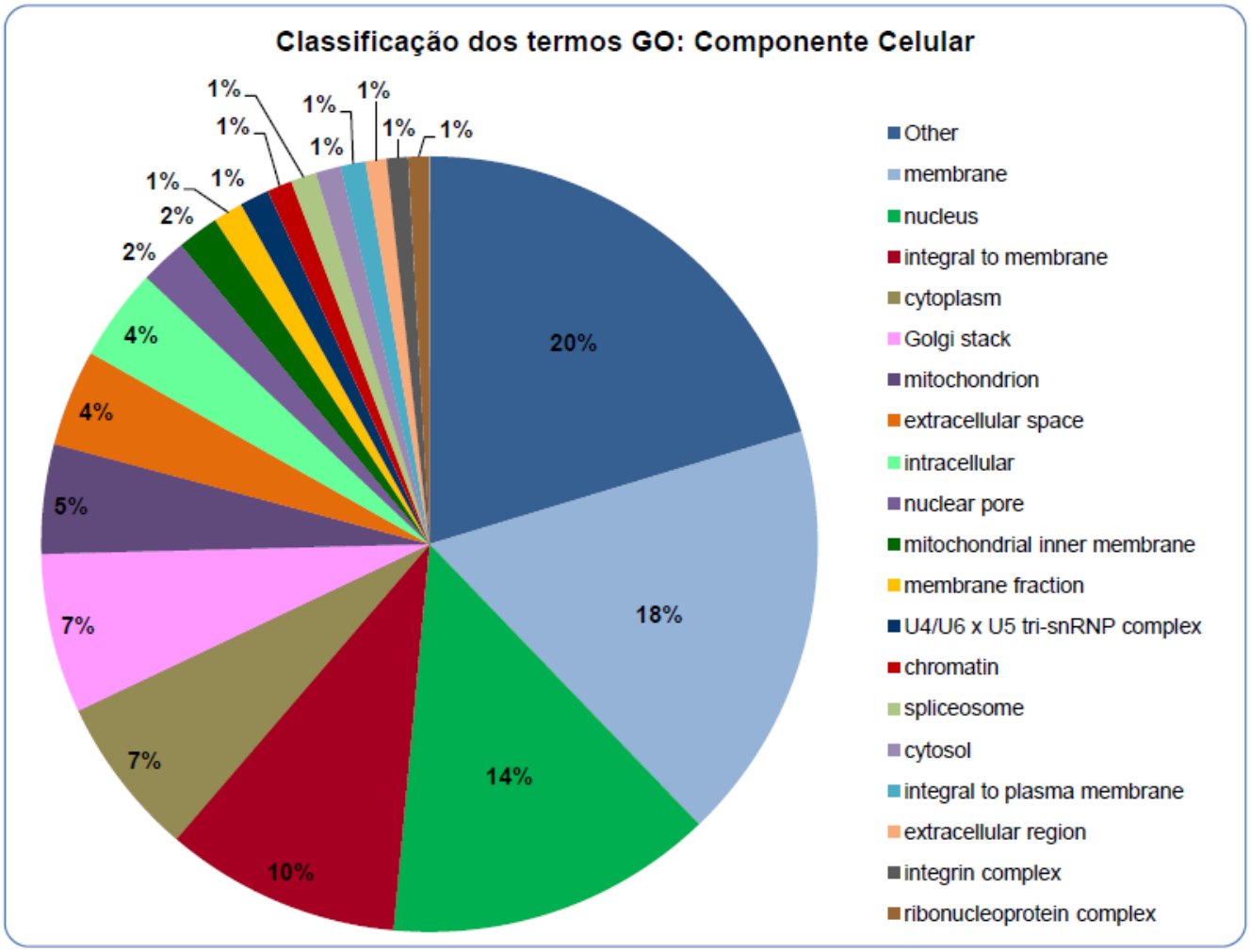


Figura 16: Agrupamento dos termos GO mais frequentes, da categoria "componente celular", transferidos para as proteínas duplicadas para as quais a função foi predita. A ordem dos termos no gráfico obedece a disposição visualizada na legenda, tendo início em "Other" e seguindo no sentido horário.

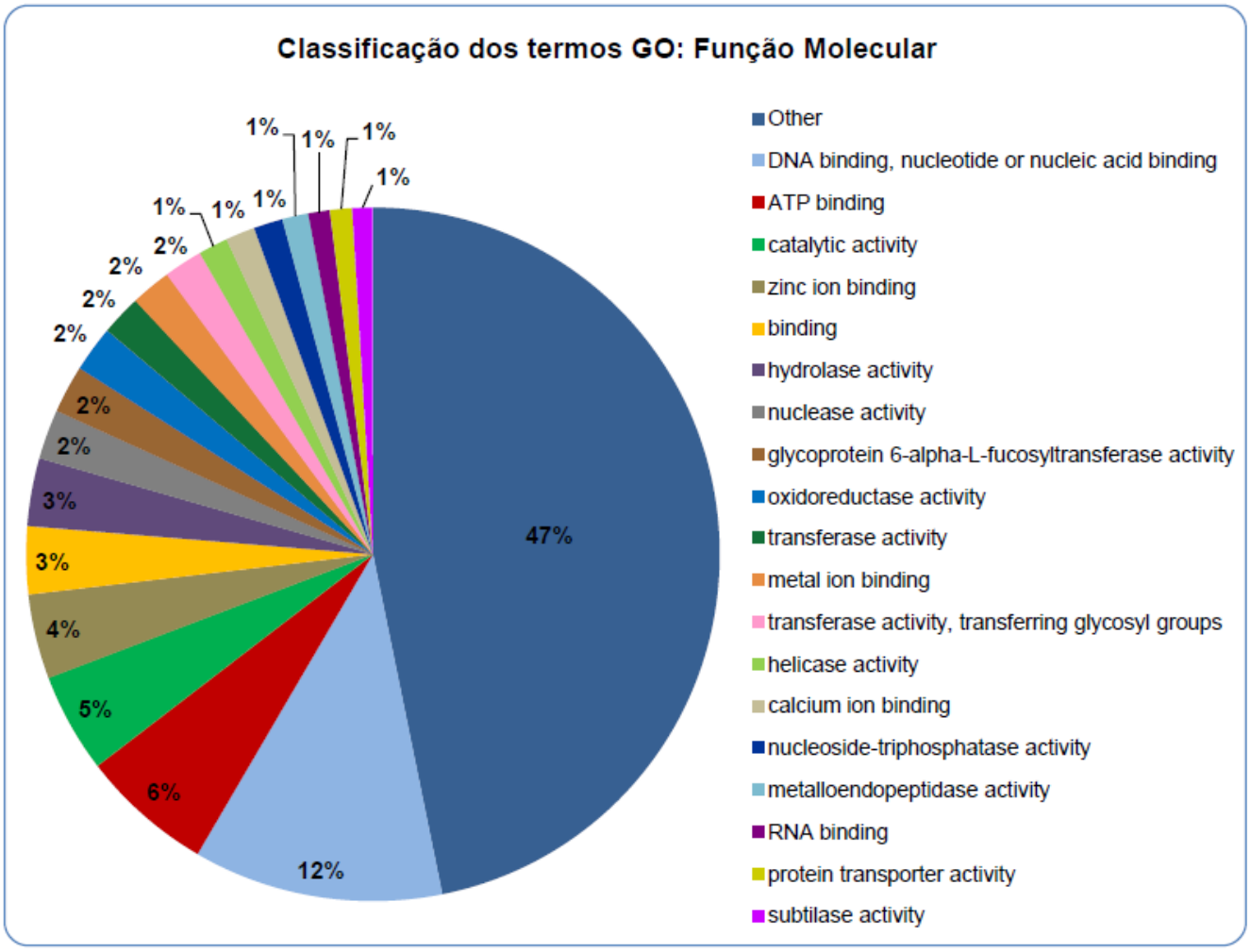


Figura 17: Agrupamento dos termos GO mais frequentes, da categoria "função molecular", transferidos para as proteínas duplicadas para as quais a função foi predita. A ordem dos termos no gráfico obedece a disposição visualizada na legenda, tendo início em "Other" e seguindo no sentido horário.

5. DISCUSSÃO

5.1. Filogenômica e anotação funcional

Ao longo dos anos, o acúmulo de dados genômicos e proteômicos tem sido muito superiores às taxas com que genes ou proteínas recebem anotação funcional, especialmente a partir de dados obtidos experimentalmente (Nahum & Pereira, 2008; Gabaldón, 2008a; Jiang, 2008; Hawkins & Kihara, 2007; Engelhardt *et al*., 2005; Sjölander, 2004; Gabaldón & Huynen, 2004). Por este motivo, a predição funcional baseada em métodos computacionais vem desempenhando um papel crucial no processo de análise de genomas recém sequenciados e na interpretação de dados experimentais produzidos em larga escala, como padrões de expressão obtidos por microarranjo ou dados de interação entre proteínas (Wu *et al.*, 2003; Hawkins & Kihara, 2007).

A análise do genoma de *S. mansoni* foi publicada recentemente. Os resultados indicam que cerca de 42% do respectivo proteoma predito permanece sem anotação funcional (Berriman *et al.*, 2009). As demais proteínas codificadas pelo genoma deste parasito tiveram a função predita através da transferência de anotação funcional dos *top hits* obtidos nas buscas por similaridade em bancos de dados de sequências. Estes resultados foram de grande importância para a comunidade científica e permitiram avanços significativos no estudo deste parasito. Nesta abordagem, o resultado das buscas é obtido através de alinhamento local, podendo ser baseado na presença de um único domínio conservado, desconsiderando a presença de domínios adicionais que estão diretamente relacionados com a função executada pela proteína. Além disso, sabe-se que genes apontados como mais similares em buscas baseadas, por exemplo, em ferramentas do pacote BLAST – *Basic Local Alignment Search Tool* (Altschul *et al.*, 1990), muitas vezes não são os mais próximos filogeneticamente (Koski & Golding, 2001). De fato, similaridade por si só não é um critério suficiente para distinguir relações de homologia (evolução divergente) de relações de homoplasia (evolução convergente) (Descorps-Declère *et al.*, 2008), motivo pelo qual a prática de predição funcional baseada em similaridade tem gerado erros sistemáticos (Brenner, 1999; Galperin & Koonin, 1998; Koski & Golding, 2001; Brown & Sjolander, 2006; Nahum & Pereira, 2008).

Vários autores reportam que a aplicação de métodos filogenéticos para realizar a predição funcional tem se mostrado uma alternativa eficaz para otimizar o processo de predição funcional e evitar a propagação de erros associados à predição baseada apenas em similaridade de sequências (Eisen, 1998; Frickey & Lupas, 2004; Brown & Sjolander, 2006; Nahum & Pereira, 2008; Huerta-Cepas *et al.*, 2010b). Esta abordagem busca

interpretar os padrões de similaridade e divergência de sequências sob uma perspectiva evolutiva (Eisen, 1998; Eisen & Wu, 2002; Levasseur *et al.*, 2008).

Tomadas em conjunto, estas considerações nos motivaram a buscar reconstruir as relações evolutivas de todas as proteínas codificadas no genoma do *S. mansoni*, conforme descrito anteriormente (Materiais e Métodos, item 3.2). O filoma de *S. mansoni*, reconstruído pela análise comparativa com os potenciais homólogos em 16 outros organismos (Quadro 1), inclui 8.818 árvores filogenéticas que abrangem 66% do proteoma predito do parasito em questão.

Para validar o *pipeline* computacional utilizado na reconstrução das árvores filogenéticas, selecionamos dois genes mitocondriais: CYTB e COXI. O genoma mitocondrial (mtDNA) de metazoários é considerado um dos melhores marcadores para a reconstrução das relações filogenéticas em diversos níveis taxonômicos, de populações a filos, sendo amplamente utilizado para a resolução de controvérsias taxonômicas (Gissi *et al*., 2008), motivo pelo qual dois genes codificados por esta organela foram escolhidos como grupo controle. Os resultados obtidos nos permite inferir que o *pipeline* é bastante robusto, permitindo a predição de potenciais ortólogos de proteínas de *S. mansoni* em outros 16 metazoários (Quadro 1). Naturalmente, o ideal seria selecionar outros marcadores (i.e. genes nucleares) para verificar o desempenho do *pipeline* ao lidar com genes nucleares de cópia única ou membros de famílias multigênicas, o que poderá ser avaliado em projetos futuros.

A análise do genoma de *S. mansoni* revelou alta incidência de *splicing* alternativo nos respectivos transcritos (Berriman *et al*., 2009). A princípio, esta ocorrência gera significativa biodiversidade. Porém, análises iniciais do filoma do parasito mostraram que produtos de *splicing* alternativo são interpretados como duplicações pelo algoritmo que realiza a predição de relações de ortologia e paralogia. Por este motivo, selecionamos o maior transcrito codificado por cada gene e os demais produtos de *splicing* alternativo foram eliminados.

Pares de proteínas que compartilham altos níveis de identidade (pelo menos 98%) também foram eliminadas, uma vez que um grande número de árvores filogenéticas (12,37% do total de árvores filogenéticas) possuíam apenas sequências altamente similares de *S. mansoni*. Inicialmente, consideramos estar diante de expansões de famílias gênicas em *S. mansoni*, porém esta hipótese não foi confirmada. Estes pares de proteínas correspondem a sequências de genes com, no mínimo, 98% de identidade e que poderiam ser fruto de eventos de duplicação muito recente, mas tais genes não foram localizados no mapa genético de *S. mansoni* (Criscione *et al.,* 2009), nem possuem domínio(s) protéico(s) rastreáveis em bancos de dados como o *Pfam* (Fiin *et al.*, 2010). Estes resultados podem

ainda indicar problemas na montagem das sequências do genoma. Em conjunto, optamos por realizar uma análise mais detalhada destas sequências no futuro.

No contexto da análise filogenética, a transferência de informação funcional é mais confiável quando baseada em relações de ortologia um-pra-um (Quadros 3 e 9), significando que há apenas um ortólogo de certo gene nas outras espécies, como foi descrito anteriormente (Huerta-Cepas *et al.*, 2010b). Quando um ou mais genes são co-ortólogos a um conjunto de genes em outro genoma, as relações de ortologia são consideradas um-pra-muitos, muitos-pra-um ou muitos-pra-muitos (Remm *et al.*, 2001). Nestes casos, duplicações que ocorreram dentro dos genomas analisados podem estar associadas a mudanças funcionais, o que afeta a confiabilidade da transferência funcional (Remm *et al.*, 2001; Huerta-Cepas *et al.*, 2010b). Entretanto, embora menos confiáveis, as predições baseadas nas relações um-pra-muitos, muitos-pra-um ou muitos-pra-muitos podem fornecer sugestões importantes na predição da atual função gênica, como por exemplo, a função de um grupo de genes co-ortólogos pode ter sido transferida para um único gene em uma relação muitos-pra-um (Huerta-Cepas *et al.*, 2010b).

O presente trabalho permitiu realizar a predição funcional de 946 proteínas (7% do proteoma predito) de *S. mansoni*, ainda sem caracterização experimental. Destas, a anotação da função, com grau de confiabilidade significativo, foi obtida para 677 (5% do proteoma predito e 12% do proteoma predito sem função conhecida), para as quais a transferência de informação funcional foi desempenhada utilizando relações de ortologia um-pra-um.

A disponibilização destas informações visa contribuir para o estudo da Biologia deste parasito, pois somente através do conhecimento do completo repertório protéico expresso em um organismo seremos capazes de compreender a complexidade dos sistemas biológicos. As demais transferências baseadas nas relações de ortologia (um-pra-muitos, muitos-pra-um e muitos-pra-muitos) permitiram sugerir a função de outras 269 proteínas (4,8% do proteoma predito sem função conhecida) e confirmar, complementar ou até mesmo alterar a anotação funcional obtida previamente para 4.561 sequências proteicas (34% do genoma predito), uma vez que a predição funcional baseada em análises filogenéticas são mais acuradas que aquelas baseadas apenas em similaridade (Eisen, 1998; Frickey & Lupas, 2004; Brown & Sjolander, 2006; Nahum & Pereira, 2008; Huerta-Cepas *et al.*, 2010b). Vale ainda ressaltar que a utilização de dados estruturais (Gerlt & Babbitt, 2000; Sjölander, 2010) e a discriminação de genes e produtos gênicos que possuem caracterização experimental aumentam consideravelmente a qualidade da predição funcional (Nahum & Pereira, 2008). Estes dados, não foram integrados às análises do presente trabalho.

5.2. Processos evolutivos e Biologia do parasito

Os parasitas do gênero *Schistosoma* foram os primeiros platelmintos a serem completamente sequenciados, logo as análises destes dados são capazes de fornecer informações importantes no que diz respeito à evolução de parasitas lofotrocozoários, superfilo que engloba os animais lofoforados e os animais que apresentam larva trocófora (moluscos, anelídeos, sipunculídeos entre outros) (Halanych *et al.*, 1995).

Em virtude da ausência de sequências similares nos bancos de dados para mais de 40% das proteínas preditas de *S. mansoni* e do alto índice de ocorrência de erros associados à predição baseada apenas em similaridade de sequências, utilizamos a análise filogenética como uma alternativa para analisar o proteoma predito deste parasito. Além de promover uma melhora significativa de 17% na predição funcional de proteínas com função desconhecida, os dados resultantes deste estudo podem prover informações importantes no que diz respeito aos mecanismos envolvidos na evolução dos genomas, genes e seus produtos e suas consequências no estudo da Biologia de Sistemas de *S. mansoni*.

Os processos de evolução molecular incluem duplicação (de genes, cromossomos, genomas), fusão e fissão gênica, embaralhamento de domínios, ganho e perda gênica, dentre outros. Eventos de duplicação gênica são os principais mecanismos de evolução molecular, fornecendo grande parte da matéria-prima a partir da qual, por divergência, a diversidade funcional evolui (Fitch, 1970; Gibson, 2009; Nahum & Pereira, 2008). Por este motivo e em função da possível relação entre os eventos de duplicação e a adaptação dos organismos a diferentes ambientes, estes processos despertam cada vez mais a atenção da comunidade científica (Wapinski *et al.*, 2007; Emes & Yang, 2008; Sun *et al.*, 2010).

Neste trabalho, foi possível identificar que aproximadamente 15% dos genes de *S. mansoni* possuem, no mínimo, um parálogo no genoma e se considerarmos as sequências eliminadas em função dos altos índices de identidade compartilhada entre elas este valor é ainda maior. Entretanto, este número é muito inferior à porcentagem de genes duplicados nos genomas de outros eucariotos que varia entre 30 e 65% (Zhang, 2003), sendo que em *C. elegans* este valor é igual a 49% (Rubin, 2000). Este fato é justificado em função do número de duplicações aqui mencionadas corresponderem, em sua grande maioria, apenas a eventos de duplicação que ocorreram após a divergência entre *S. mansoni* e os demais táxons selecionados (Quadro 1). A identificação de todas as duplicações não foi realizada em função de uma limitação metodológica, uma vez que o programa utilizado inicia a análise na proteína “semente” e termina a varredura na raiz da árvore. Analisando a árvore filogenética da proteína “semente” Sch0013135 (Figura 14) é possível compreender como o programa atua. O identificador Sch0009447 não é contabilizado como parálogo em função

deste estar localizado na porção inferior à raiz da árvore. Como o programa interpreta a raiz como sendo o ponto final da análise, ao identificá-la, o programa finaliza a busca e a outra partição da árvore não é examinada.

Como observado previamente por outros autores, identificamos dentre as duplicações de *S. mansoni* genes parálogos que codificam leishmanolisinas (metaloprotease relacionada à invasão do parasito no corpo do hospedeiro definitivo), tetraspaninas (proteínas de membrana) e fucosiltransferases (enzimas envolvidas na produção de estruturas imunomodulatórias durante a infecção) (Berriman *et al*., 2009; Marques *et al.*, 2001; Fitzpatrick *et al.*, 2009). Estas proteínas desempenham funções críticas na interface parasito-hospedeiro, motivo pelo qual são amplamente estudadas como alvos para drogas e vacinas contra a esquistossomose (Tran *et al.*, 2006; McManus *et al.*, 2008; Fitzpatrick *et al.*, 2009).

Abordando outro mecanismo evolutivo, Yu e colaboradores (2008) analisaram a relação parasito-hospedeiro entre *H. sapiens* e *S. japonicum* em um estudo que revelou altos índices de similaridade entre algumas sequências destes organismos, que os autores interpretaram como sendo reflexo da ocorrência de transferência horizontal de genes entre o hospedeiro vertebrado e o parasito. Outros autores também reportaram evidências de transferência horizontal de sequências gênicas de *H. sapiens* para parasitos do gênero *Schistosoma* (Imase *et al*., 2004; Okada *et al.*, 2006; DeMarco *et al.*, 2007). Dentre as árvores filogenéticas resultantes do presente trabalho, oito possuem apenas sequências de *S. mansoni* e *H. sapiens* e em outros casos, embora existam sequências de outros organismos na árvore, os identificadores do parasito e do hospedeiro definitivo estão presentes de forma isolada no mesmo clado, conforme apresentado anteriormente (Resultados, item 4.4). A princípio, estes dados poderiam corroborar a hipótese de eventos de transferência horizontal como uma explicação plausível para a co-ocorrência de clados onde estão presentes apenas sequências destes dois organismos, embora não exista proximidade filogenética entre estas espécies. Entretanto, este não é o único mecanismo capaz de explicar o fato observado, uma vez que mecanismos de coevolução parasito-hospedeiro também podem resultar nos padrões filogenéticos encontrados. Além disso, observamos em nosso trabalho outras instâncias nas quais proteínas de *S. mansoni* estão representadas em árvores filogenéticas juntamente com um único organismo dentre os selecionados para análise, como por exemplo, árvores com proteínas de *S. mansoni* e *C. elegans*.

6. CONCLUSÕES

Através de estudos de filogenia molecular podemos compreender mais amplamente os dados de sequências genômicas e as informações referentes à estrutura e função das proteínas codificadas por estes genomas que vem se acumulando rapidamente (Whelan *et al.*, 2001). A análise filogenômica aqui apresentada gerou um total de 53.189 transferências de anotação funcional para 5.507 sequências de aminoácidos de *S. mansoni*. A predição funcional de 946 proteínas desconhecidas até o momento, bem como a confirmação, complementação ou até mesmo alteração da função de outras 4.561 sequências, aprimoraram a anotação do proteoma predito deste parasito, disponibilizando uma base de dados mais robusta para a comunidade científica. Além de promover uma melhora significativa na predição funcional do proteoma de *S. mansoni*, este estudo provê informações importantes sobre a evolução do genoma desta espécie, como a identificação de duplicações gênicas que podem estar relacionadas a especificidades morfológicas ou fisiológicas deste organismo.

7. LIMITAÇÕES DO ESTUDO

Em síntese, os resultados deste trabalho corroboram a literatura que menciona o ganho significativo na acurácia da predição funcional de genes e produtos gênicos baseada em métodos filogenéticos (Eisen, 1998; Sjölander, 2004; Brown & Sjölander, 2006; Fuellen, 2008; Nahum & Pereira, 2008) e o potencial das análises comparativas de genomas como uma ferramenta para a exploração dos processos evolutivos envolvidos na evolução dos organismos vivos (Levasseur *et al.*, 2008; Boussau & Daubin, 2010). Entretanto, para que a comunidade científica possa se beneficiar dos resultados da predição funcional fruto deste estudo, é necessário conhecer e compreender as limitações envolvidas na aplicação da abordagem filogenômica à anotação funcional no contexto aqui apresentado. Estas pesquisas são projetadas para fornecer à comunidade científica informações no que diz respeito à função mais provável a ser executada pelas proteínas, mas não substitui de forma alguma a confirmação experimental, uma vez que os métodos filogenéticos, assim como os demais, possuem limitações e estão sujeitos a erros, especialmente se não forem aplicados corretamente. Em especial, é preciso lembrar que ortologia é um conceito estritamente evolutivo, certamente relacionado, mas não baseado na funcionalidade das sequências envolvidas (Gabaldón, 2008b). Na maioria das vezes, a diversidade funcional está associada à divergência significante no nível da sequência, mas altos níveis de identidade

nem sempre garantem que duas ou mais proteínas executem a mesma função, uma vez que alterações sutis em sítios ativos são capazes de alterar completamente a função desempenhada pela proteína (Gerlt & Babbitt, 2000). Portanto, as predições funcionais obtidas neste trabalho para as proteínas de *S. mansoni* devem ser utilizadas como ponto de partida para projetos futuros, priorizando a escolha de determinados genes ou proteínas em novos estudos experimentais.

8. PERSPECTIVAS

Como perspectivas para este estudo, sugerimos uma análise detalhada das relações evolutivas entre *S. mansoni* e os outros metazoários. Seria certamente muito interessante prosseguir as investigações acerca dos eventos de duplicação, perda e/ou ganho gênico e seu papel na evolução do parasitismo. Buscando-se compreender melhor os mecanismos envolvidos no processo evolutivo do genoma e proteoma de *S. mansoni*, estaremos contribuindo para o estudo da Biologia de Sistemas deste parasito. As duplicações evidenciadas neste estudo, principalmente aquelas que ocorreram após a divergência entre *S. mansoni* e os demais táxons, podem estar relacionadas a especificidades morfológicas ou fisiológicas deste parasito, constituindo uma vasta fonte de informações para estudos futuros com o objetivo de identificar novos alvos para o desenvolvimento de drogas, vacinas e diagnósticos contra a esquistossomose.

9. REFERÊNCIAS BIBLIOGRÁFICAS

Adachi J, Hasegawa M. Model of amino acid substitution in proteins encoded by mitochondrial DNA. **Journal of Molecular Evolution**, v. 42, p. 459-68, 1996.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC. The genome sequence of *Drosophila melanogaster*. **Science**, v. 287, p. 2185-95, 2000.

Akaike H. Information theory and extension of the maximum likelihood principle. **Proceedings of the 2nd international symposium on information theory**. Budapest, Hungary, p. 267-281, 1973.

Alam K, Maheshwari V, Jain A, Siddiqui FA, Haq ME, Prasad S, Hasan AS. Schistosomiasis: A Case Series, With Review Of Literature. **The Internet Journal of Infectious Diseases**, v. 7, n. 1, 2009.

Almeida-King JP, Donaldson S, Laird GK, LloydDM, Sehra HK, Collins JE, Howe K, Reimholz B, Torrance J, Trevanion S, Stemple D, Barroso I, Gilbert JGR, Griffiths E, Loveland JE, Storey R, Harrow JL, Hubbard T. Update On The Zebrafish Genome Project. **3rd International Biocuration Conference**, 2009.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, p. 403-10, 1990.

Anisimova M, Gascuel O. Approximate likelihoodratio test for branches: a fast, accurate, and powerful alternative. **Systematic Biology**, v. 55, p. 539–552, 2006.

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, v. 408, p. 796-815, 2000.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nature Genetics**, v. 25, p. 25-9, 2000.

Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD, Aslett MA, Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA, Delcher A, DeMarco R, Djikeng A, Eyre T, Gamble JA, Ghedin E, Gu Y, Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, Lacerda D, Macedo CD, McVeigh P, Ning Z, Oliveira G, Overington JP, Parkhill J, Pertea M, Pierce RJ, Protasio AV, Quail MA, Rajandream MA, Rogers J, Sajid M, Salzberg SL, Stanke M, Tivey AR, White O, Williams DL, Wortman J, Wu W, Zamanian M, Zerlotini A, Fraser-Liggett CM, Barrell BG, El-Sayed NM. The genome of the blood fluke *Schistosoma mansoni*. **Nature**, v.460, p. 352–358, 2009.

Bieberich AA, Minchella DJ. Mitochondrial inheritance in Schistosoma mansoni: mitochondrial variable number tandem repeat mutation produces noise on top of the signal. The Journal of Parasitology, v. 87, p. 1011-5, 2001.

Botros SS, Bennett JL. Praziquantel resistance. **Expert Opinion on Drug Discovery**, v. 2, p. S35-S40, 2007.

Boussau B, Daubin V. Genomes as documents of evolutionary history. **Trends in Ecology and Evolution**, v. 25, p. 224-32, 2010.

Brenner SE. Errors in genome annotation. **Trends in Genetics**, v.15, p. 132-3, 1999.

Brown D, Sjolander K. Functional classification using phylogenomics inference. **PLoS Computational Biology**, v. 2, e. 77, 2006.

Bruun B, Aagaard-Hansen J. The social context of schistosomiasis and its control: an introduction and annotated bibliography. Tropical Disease Research (TDR), 2008. Disponível em:

<http://apps.who.int/tdr/svc/publications/tdr-research-publications/social-context-schistosomiasis>. Acesso em 21 de jan. 2010.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. **Bioinformatics**, v. 25, p. 1972-3, 2009.

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D. Genetic and physical maps of *Saccharomyces cerevisiae*. **Nature**, v. 29, p. 67-73, 1997.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, Hlavina W, Kapustin Y, Meric P, Maglott D, Birtle Z, Marques AC, Graves T, Zhou S, Teague B, Potamousis K, Churas C, Place M, Herschleb J, Runnheim R, Forrest D, Amos-Landgraf J, Schwartz DC, Cheng Z, Lindblad-Toh K, Eichler EE, Ponting CP; Mouse Genome Sequencing Consortium. Lineage-specific biology revealed by a finished genome assembly of the mouse. **PLoS Biology**, v. 7, e. 1000112, 2009.

Criscione CD, Valentim CL, Hirai H, LoVerde PT, Anderson TJ. Genomic linkage map of the human blood fluke Schistosoma mansoni. **Genome Biology**, v. 10, R71, 2009.

DbVisualizer: the universal database tool. Disponível em: <http://www.dbvis.com>. Acesso em: 15 set. 2010.

Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, Harafuji N, Hastings KE, Ho I, Hotta K, Huang W, Kawashima T, Lemaire P, Martinez D, Meinertzhagen IA, Necula S, Nonaka M, Putnam N, Rash S, Saiga H, Satake M, Terry A, Yamada L, Wang HG, Awazu S, Azumi K, Boore J, Branno M, Chin-Bow S, DeSantis R, Doyle S, Francino P, Keys DN, Haga S, Hayashi H, Hino K, Imai KS, Inaba K, Kano S, Kobayashi K, Kobayashi M, Lee BI, Makabe KW, Manohar C, Matassi G, Medina M, Mochizuki Y, Mount S, Morishita T, Miura S, Nakayama A, Nishizaka S, Nomoto H, Ohta F, Oishi K, Rigoutsos I, Sano M, Sasaki A, Sasakura Y, Shoguchi E, Shin-i T, Spagnuolo A, Stainier D, Suzuki MM, Tassy O, Takatori N, Tokuoka M, Yagi K, Yoshizaki F, Wada S, Zhang C, Hyatt PD, Larimer F, Detter C, Doggett N, Glavina T, Hawkins T, Richardson P, Lucas S, Kohara Y, Levine M, Satoh N, Rokhsar DS. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. **Science**, v. 298, p. 2157-67, 2002.

Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. **Nature Review Genetics**, v. 6, p. 361-75, 2005.

DeMarco R, Mathieson W, Dillon GP, Wilson RA. Schistosome albumin is of host, not parasite, origin. **International Journal for Parasitology**, v. 37, p. 1201-8, 2007.

Descorps-Declère S, Lemoine F, Sculo Q, Lespinet O, Labedan B. The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species. **Biochimie**, v. 90, p. 595-608, 2008.

Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. **BMC Bioinformatics**, v. 5, n. 113, 2004.

Eisen, JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. **Genome Research**, v. 8, p. 163-167, 1998.

Eisen JA, Wu M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. **Theoretical Population Biology**, v. 61, p. 481–487, 2002.

Emes RD, Yang Z. Duplicated paralogous genes subject to positive selection in the genome of *Trypanosoma brucei*. **PLoS ONE**, v. 3, e. 2295, 2008.

Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. **PLoS Computational Biology**, v.1, e.45, 2005.

Engels D, Chitsulo L, Montresor A, Savioli L. The global epidemiological situation of schistosomiasis and new approaches to control and research. **Acta Tropica**, v. 82, p. 139–46, 2002.

Engels D, Chitsulo L. Schistosomiasis. In: Crompton DWT, Montresor A, Nesheim MC, Savioli L, Controlling disease due to helminth infections. World Health Organization. Geneva, Switzerland, 2003. Disponível em <http://www.who.int/wormcontrol/documents/en/Controlling%20Helminths.pdf>. Acesso em: 01 de maio de 2010.

ETE: a python Environment for Tree Exploration. Disponível em: <http://ete.cgenomics.org>. Acesso em: 26 nov. 2010.

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach, **Journal of Molecular Evolution**, v. 17, p. 368-76, 1981.

Ferreira JE, Takai OK. Understanding Database Design. In: Gruber A, Durham AM, Huynh C, del Portillo HA (eds), Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach. National Center for Biotechnology Information (US), 2008. cap. A02.

FigTree: graphical viewer of phylogenetic trees. Disponível em: <http://tree.bio.ed.ac.uk/software/figtree/>. Acesso em: 22 fev. 2010.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A. The Pfam protein families database. **Nucleic Acids Research**, v. 38, p. D211-D222, 2010.

Frickey T, Lupas AN. PhyloGenie: automated phylome generation and analysis. **Nucleic Acids Research**, v. 32, p. 5231-8, 2004.

Fitch,WM. Distinguishing homologous from analogous proteins. **Systematic. Zoology**, v.19, p. 99–113, 1970.

Fitzpatrick JM, Peak E, Perally S, Chalmers IW, Barrett J, Yoshino TP, Ivens AC, Hoffmann KF. Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses. **PLoS Neglected Tropical Diseases**, v. 3, e. 543, 2009.

Frickey T, Lupas AN. PhyloGenie: automated phylome generation and analysis. **Nucleic Acids Research**, v. 32, p. 5231-8, 2004.

Fuellen G. Homology and phylogeny and their automated inference. **Naturwissenschaften**, v. 95, p. 469-81, 2008.

Gabaldón T. Evolution of proteins and proteomes: a phylogenetics approach. **Evolutionary Bioinformatics**, v. 1, p. 51–61, 2005.

Gabaldón T. Comparative Genomics-Based Prediction of Protein Function. **Genomics Protocols**, v. 439, p. 387-401, 2008a.

Gabaldón T. Large-scale assignment of orthology: back to phylogenetics? **Genome Biology**, v. 9, n. 235, 2008b.

Gabaldón T, Huynen MA. Prediction of protein function and pathways in the genome era. **Cellular and Molecular Life Sciences**, v.61, p. 930-44, 2004.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA,

Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B. The genome sequence of the filamentous fungus Neurospora crassa. **Nature**, v. 422, p. 859-68, 2003.

Galperin MY, Koonin EV. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. **In Silico Biology**, v.1, p. 55-67, 1998.

Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. **Nature Biotechnology**, v. 18, p. 609-13, 2000.

Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. **Molecular Biology and Evolution**, v. 14, p. 685-95, 1997.

Gering EJ, Opazo JC, Storz JF. Molecular evolution of cytochrome b in high- and low-altitude deer mice (genus Peromyscus). **Heredity**, v. 102, p. 226-235, 2009.

Gerlt JA, Babbitt PC. Can sequence determine function? **Genome Biology**, v. 1, p. 0005.1–0005.10, 2000.

Gibson TA, Goldberg DS. Questioning the ubiquity of neofunctionalization. **PLoS Computational Biology**, v. 5, e.1000252, 2009.

Gissi C, Iannelli F, Pesole G. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. **Heredity**, v. 101, p. 301-20, 2008.

Gryseels B, Polman K, Clerinx J, Kestens L. Human schistosomiasis. **Lancet,** v. 368, p. 1106-18, 2006.

Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. **Systematic Biology**, v. 52, p. 696-704, 2003.

Han ZG, Brindley PJ, Wang SY, Chen Z. Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction. **Annual Review of Genomics and Human Genetics**, v. 10, p. 211-40, 2009.

Halanych KM, Bacheller JD, Aguinaldo AMA, Liva SM, Hillis DM, Lake JA. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. **Science**, v. 267, p. 1641-1643, 1995.

Hawkins T, Kihara D. Function prediction of uncharacterized proteins. Journal **of Bioinformatics and Computacional Biology**, v.5, p. 1-30, 2007.

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 89, p. 10915–9, 1992.

Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. **Nature Reviews Genetics**, v. 4, p. 275-284, 2003.

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL. The genome sequence of the malaria mosquito *Anopheles gambiae*. **Science**, v. 298, p. 129-49, 2002.

Hood L. Systems biology: integrating technology, biology, and computation. **Mechanisms of Ageing and Development**, v. 124, p. 9-16, 2003.

Hotez PJ, Kamath A. Neglected Tropical Diseases in Sub-Saharan Africa: Review of Their Prevalence, Distribution, and Disease Burden. **PLoS Neglected Tropical Diseases**, v. 3, e.412, 2009.

Hu M, Gasser RB. Mitochondrial genomes of parasitic nematodes - progress and perspectives. **Trends in Parasitology**, v. 22, p. 78-84, 2006.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. The human phylome. **Genome Biology**, v. 8, R109, 2007.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. PhylomeDB: a database for genome-wide collections of gene phylogenies. **Nucleic Acids Research**, v. 36, p. D491-D496, 2008.

Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. **BMC Bioinformatics**, v. 11, n. 24, 2010a.

Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldón T. The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes. **Insect Molecular Biology**, v. 19, p. 13-21, 2010b.

Imase A, Ohmae H, Iwamura Y, Kirinoki M, Matsuda H. A comparative study on mouse MHC class I sequences detected in Schistosoma japonicum recovered from BALB/c (H-2d) and C57BL/6 (H-2b) mice. **The Southeast Asian Journal of Tropical Medicine and Public Health**, v. 35, p. 10-8, 2004.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. **Nature**, v. 431, p. 931-45, 2004.

Jiang Z. Protein Function Predictions Based on the Phylogenetic Profile Method. **Critical Reviews in Biotechnology**, v. 28, p. 233–238, 2008.

Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. **Computer Applications in the Biosciences**, v.8, p. 275-282, 1992.

Kelchner SA, Thomas MA. Model use in phylogenetics: nine key questions. **Trends in Ecology & Evolution**, v. 22, p. 87-94, 2007.

Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annual Review of Genetics, v. 39, p. 309-38, 2005.

Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. **Journal of Molecular Evolution**, v. 52, p. 540-2, 2001.

Kranthi S, Kranthi KR, Bharose AA, Syed SN, Dhawad CS, Wadaskar RM, Behere GT, Patil EK. Cytochrome oxidase I sequence of *Helicoverpa* (Noctuidae: Lepidoptera) species in India - Its utility as a molecular tool. **Indian Journal of Biotechnology**, v. 5, p. 195-199, 2006.

Le TH, Blair D, McManus DP. Mitochondrial genomes of parasitic flatworms. **Trends in Parasitology**, v. 18, p. 206-13, 2002.

Le TH, Humair PF, Blair D, Agatsuma T, Littlewood DT, McManus DP. Mitochondrial gene content, arrangement and composition compared in African and Asian schistosomes. **Molecular and Biochemical Parasitology**, v. 117, p. 61-71, 2001.

Levasseur A, Pontarotti P, Poch O, Thompson JD. Strategies for Reliable Exploitation of Evolutionary Concepts in High Throughput Biology. **Evolutionary Bioinformatics**, v. 4, p. 121–137, 2008.

Lewis SE. Gene Ontology: looking backwards and forwards. **Genome Biology**, v. 6, n. 103, 2005.

Liang YS, Dai JR, Zhu YC, Coles GC, Doenhoff MJ. Genetic analysis of praziquantel resistance in Schistosoma mansoni. **The Southeast Asian Journal of Tropical Medicine and Public Health**, v. 34, p. 274-80, 2003.

Liò P, Goldman N. Models of molecular evolution and phylogeny. **Genome Research**, v. 8, p. 1233-44, 1998.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJ, van Oudenaarden A, Barton DB, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ. Population genomics of domestic and wild yeasts. **Nature**, v. 458, p. 337-41, 2009.

LoVerde PT, Hirai H, Merrick JM, Lee NH, El-Sayed N. *Schistosoma mansoni* genome project: an update. **Parasitology International**, v. 53, p. 183-92, 2004.

Marques Jr ETA, Ichikawa Y, Strand M, August JT, Hart GW, Schnaar RL. Fucosyltransferases in Schistosoma mansoni development. **Glycobiology**, v. 11, p. 249-59, 2001.

McManus DP, Loukas A. Current status of vaccines for schistosomiasis. **Clinical Microbiology Reviews**, v. 21, p. 225-42, 2008.

Medina M. Genomes, phylogeny, and evolutionary systems biology. **Proceedings of the National Academy of Sciences**, v. 102, p. 6630-6635, 2005.

Melman SD, Steinauer ML, Cunningham C, Kubatko LS, Mwangi IN, Wynn NB, Mutuku MW, Karanja DM, Colley DG, Black CL, Secor WE, Mkoji GM, Loker ES. Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of Schistosoma mansoni. **PLoS Neglected Tropical Disease**, v. 3, e. 504, 2009.

Melo AL, Coelho PMZ. *Schistosoma mansoni* e a doença. In: Neves DP, Melo AL, Linardi PM, Vitor RWA (eds). Parasitologia Humana. 11 ed. Belo Horizonte, 2005, cap. 22, p. 193-212.

Microscopia eletrônica de um casal de parasitos da espécie *S. mansoni*. Disponível em <http://emtrix.dbs.umt.edu/imagery/imagery.htm>. Acesso em: 30 mar. 2010.

Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H, Namiki N, Kitagawa M, Yamashita H, Yasukochi Y, Kadono-Okuda K, Yamamoto K, Ajimura M, Ravikumar G, Shimomura M, Nagamura Y, Shin-I T, Abe H, Shimada T, Morishita S, Sasaki T. The genome sequence of silkworm, *Bombyx mori*. **DNA Research**, v. 11, p. 27-35, 2004.

Muller T, Vingron M. Modeling amino acid replacement. **Journal of Computational Biology**, v. 7, p. 761-776, 2000.

Nahum LA, Pereira SL. Phylogenomics, Protein Family Evolution, and the Tree of Life: An Integrated Approach between Molecular Evolution and Computational Intelligence. In: Smolinski TG, Milanova MG, Hassanien A-E (eds), Studies in Computational Intelligence (SCI) 122. Springer-Verlag Berlin Heidelberg, 2008. cap. 11, p. 259–279.

NCBI: The Genome Database. Disponível em: <http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome>. Acesso em: 19 mar. 2010.

O'Brien EA, Zhang Y, Wang E, Marie V, Badejoko W, Lang BF, Burger G. GOBASE: an organelle genome database. **Nucleic Acids Research**, v. 37, p. D946-50, 2009.

Okada A, Imase A, Matsuda H, Ohmae H, Hata H, Iwamura Y. Heterogeneity of class I and class II MHC sequences in Schistosoma mansoni. **Parasitolology Research**, v. 99, p. 21-7, 2006.

Page RDM, Holmes EC. Molecular Evolution. A phylogenetic approach. Blackwell Science Ltd. Oxford, UK, 1998, 346 p.

PhylomeDB: a public database for complete collections of gene phylogenies (phylomes). Disponível em: <phylomedb.org>. Acesso em: 28 abr. 2010.

Pica-Mattoccia L, Cioli D. Sex- and stage-related sensitivity of Schistosoma mansoni to in vivo and in vitro praziquantel treatment. **International Journal for Parasitology**, v. 34, p. 527-33, 2004.

Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. **Systematics Biology**, v.53, p.793-808, 2004.

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS. Sea anemone genome reveals ancestral eumetazoan gener epertoire and genomic organization. **Science**, v. 317, p. 86-94, 2007.

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutiérrez EL, Dubchak I, Garcia-Fernàndez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PW, Satoh N, Rokhsar DS. The amphioxus genome and the evolution of the chordate karyotype. **Nature**, v. 453, p. 1064-71, 2008.

Rambaut, A. Tree Figure Drawing Tool, Version 1.3.1. Institute of Evolutionary Biology, University of Edinburgh. Disponível em: <http://tree.bio.ed.ac.uk/software/figtree>. Acesso em: 3 jan. 2010.

Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. **Journal of Molecular Biology**, v. 314, p. 1041-52, 2001.

Ridley M. Evolução. 3. ed. Porto Alegre: Editora Artmed, 2006. cap. 15, p. 447-495.

Ross AG, Bartley PB, Sleigh AC, Olds GR, Li Y, Williams GM, McManus DP. Schistosomiasis. **The New England Journal of Medicine**, v. 346, p. 1212-20, 2002.

Rubin GM. Comparative genomics of the eukaryotes. **Science**, v. 287, p. 2204–2215, 2000.

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, v. 4, p. 406-25, 1987.

Saribaş AS, Valkova-Valchanova M, Tokito MK, Zhang Z, Berry EA, Daldal F. Interactions between the cytochrome b, cytochrome c1, and Fe-S protein subunits at the ubihydroquinone oxidation site of the bc1 complex of Rhodobacter capsulatus. **Biochemistry**, v. 37, p. 8105-14, 1998.

Schneider, H. Métodos de análise filogenética. 3. ed. Ribeirão Preto: Holos Editora, 2007. cap. 6, p. 56-71.

SchistoDB: *Schistosoma mansoni* Database. Disponível em: <schistodb.net>. Acesso em: 24 mar. 2010.

Sea Urchin Genome Sequencing Consortium, Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, Angerer RC, Angerer LM, Arnone MI, Burgess DR, Burke RD, Coffman JA, Dean M, Elphick MR, Ettensohn CA, Foltz KR, Hamdoun A, Hynes RO, Klein WH, Marzluff W, McClay DR, Morris RL, Mushegian A, Rast JP, Smith LC, Thorndyke MC, Vacquier VD, Wessel GM, Wray G, Zhang L, Elsik CG, Ermolaeva O, Hlavina W, Hofmann G, Kitts P, Landrum MJ, Mackey AJ, Maglott D, Panopoulou G, Poustka AJ, Pruitt K, Sapojnikov V, Song X, Souvorov A, Solovyev V, Wei Z, Whittaker CA, Worley K, Durbin KJ, Shen Y, Fedrigo O, Garfield D, Haygood R, Primus A, Satija R,

Severson T, Gonzalez-Garay ML, Jackson AR, Milosavljevic A, Tong M, Killian CE, Livingston BT, Wilt FH, Adams N, Bellé R, Carbonneau S, Cheung R, Cormier P, Cosson B, Croce J, Fernandez-Guerra A, Genevière AM, Goel M, Kelkar H, Morales J, Mulner-Lorillon O, Robertson AJ, Goldstone JV, Cole B, Epel D, Gold B, Hahn ME, Howard-Ashby M, Scally M, Stegeman JJ, Allgood EL, Cool J, Judkins KM, McCafferty SS, Musante AM, Obar RA, Rawson AP, Rossetti BJ, Gibbons IR, Hoffman MP, Leone A, Istrail S, Materna SC, Samanta MP, Stolc V, Tongprasit W, Tu Q, Bergeron KF, Brandhorst BP, Whittle J, Berney K, Bottjer DJ, Calestani C, Peterson K, Chow E, Yuan QA, Elhaik E, Graur D, Reese JT, Bosdet I, Heesun S, Marra MA, Schein J, Anderson MK, Brockton V, Buckley KM, Cohen AH, Fugmann SD, Hibino T, Loza-Coll M, Majeske AJ, Messier C, Nair SV, Pancer Z, Terwilliger DP, Agca C, Arboleda E, Chen N, Churcher AM, Hallböök F, Humphrey GW, Idris MM, Kiyama T, Liang S, Mellott D, Mu X, Murray G, Olinski RP, Raible F, Rowe M, Taylor JS, Tessmar-Raible K, Wang D, Wilson KH, Yaguchi S, Gaasterland T, Galindo BE, Gunaratne HJ, Juliano C, Kinukawa M, Moy GW, Neill AT, Nomura M, Raisch M, Reade A, Roux MM, Song JL, Su YH, Townley IK, Voronina E, Wong JL, Amore G, Branno M, Brown ER, Cavalieri V, Duboc V, Duloquin L, Flytzanis C, Gache C, Lapraz F, Lepage T, Locascio A, Martinez P, Matassi G, Matranga V, Range R, Rizzo F, Röttinger E, Beane W, Bradham C, Byrum C, Glenn T, Hussain S, Manning G, Miranda E, Thomason R, Walton K, Wikramanayke A, Wu SY, Xu R, Brown CT, Chen L, Gray RF, Lee PY, Nam J, Oliveri P, Smith J, Muzny D, Bell S, Chacko J, Cree A, Curry S, Davis C, Dinh H, Dugan-Rocha S, Fowler J, Gill R, Hamilton C, Hernandez J, Hines S, Hume J, Jackson L, Jolivet A, Kovar C, Lee S, Lewis L, Miner G, Morgan M, Nazareth LV, Okwuonu G, Parker D, Pu LL, Thorn R, Wright R. The genome of the sea urchin Strongylocentrotus purpuratus. **Science**, v. 314, p. 941-52, 2006.

Sharakhova MV, Hammond MP, Lobo NF, Krzywinski J, Unger MF, Hillenmeyer ME, Bruggner RV, Birney E, Collins FH. Update of the Anopheles gambiae PEST genome assembly. **Genome Biology**, v. 8, n. R5, 2007.

Short RB, Menzel MY, Pathak S. Somatic chromosomes of *Schistosoma mansoni*. **The Journal of Parasitology**, v. 65, p. 471-3, 1979.

Sicheritz-Pontén T, Andersson SG. A phylogenomic approach to microbial evolution. **Nucleic Acids Research**, v. 29, p. 545-52, 2001.

Sjölander K. Phylogenomic inference of protein molecular function: advances and challenges. **Bioinformatics**, v. 20, p. 170-179, 2004.

Sjölander K. Getting started in structural phylogenomics. **PLoS Computational Biology**, v.6, e. 1000621, 2010.

Smith TF, Waterman MS. Identification of common molecular subsequences. **Journal of Molecular Biology**, v. 147, p. 195-197,1981.

Snel B, Huynen MA, Dutilh BE. Genome trees and the nature of genome evolution. **Annual Review of Microbiology**, v. 59, p. 191-209, 2005.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DH, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R, Waterston RH. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. **PLoS Biology**, v. 1, e. 45, 2003.

Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. **The Lancet Infectious Diseases**, v.6, p. 411-25, 2006.

Studier JA, Keppler KJ. A note on the neighbor-joining algorithm of Saitou and Nei. **Molecular Biology and Evolution**, v. 5, p. 729-31, 1988.

Sun J, Jiang H, Flores R, Wen J. Gene duplication in the genome of parasitic *Giardia lamblia*. **BMC Evolutionary Biology**, v. 10, p. 2-8, 2010.

Telford MJ. Animal phylogeny. **Current Biology**, v. 16, p. R981-R985, 2006.

The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. **Science**, v. 282, p. 2012-8, 1998.

Tran MH, Pearson MS, Bethony JM, Smyth DJ, Jones MK, Duke M, Don TA, McManus DP, Correa-Oliveira R, Loukas A. Tetraspanins on the surface of Schistosoma mansoni are protective antigens against schistosomiasis. **Nature Medicine**, v. 12, p. 835-40, 2006.

Tropical Disease Research (TDR): progress 2005-2006. Eighteenth programme report, 2007. Disponível em: <http://apps.who.int/tdr/svc/publications/about-tdr/progress-reports/progress-report-05-06>. Acesso em 21 de jan. 2010.

The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. **Nucleic Acids Research**, v. 38, p. D142-D148, 2010.

UniProt: Universal Protein Resource. Disponível em: <www.uniprot.org>. Acesso em: 13 jan. 2010.

Ustilago maydis Sequencing Project. Broad Institute of MIT and Harvard. Disponível em: <http://www.broad.mit.edu>. Acesso em: 01 jun. 2006.

Vàlles Y, Boore JL. Lophotrochozoan mitochondrial genomes. **Integrative and Comparative Biology**, v. 46, p. 544-557, 2006.

van der Werf MJ, de Vlas SJ, Brooker S, Looman CW, Nagelkerke NJ, Habbema JD, Engels D. Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa. **Acta Tropica**, v. 86, p. 125-39, 2003.

Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. **Nature**, v. 449, p. 54-64, 2007.

Webster JP, Oliviera G, Rollinson D, Gower CM. Schistosome genomes: a wealth of information. **Trends in Parasitology**, v. 26, p. 103-6, 2010.

Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. **Molecular Biology and Evolution**, v. 18, p. 691-699, 2001.

Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. **Trends in Genetics**, v. 17, p. 262-72, 2001.

Wu CH, Huang H, Yeh LS, Barker WC. Protein family classification and functional annotation. **Computational Biology and Chemistry**, v. 27, p. 37-47, 2003.

Yu F, Li Y, Liu L, Li Y. Comparative genomics of human-like *Schistosoma japonicum* genes indicates a putative mechanism for host-parasite relationship. **Genomics**, v. 91, p. 152-7, 2008.

Zerlotini A, Heiges M, Wang H, Moraes RL, Dominitini AJ, Ruiz JC, Kissinger JC, Oliveira G. SchistoDB: a Schistosoma mansoni genome resource. **Nucleic Acids Research**, v. 37, p. D579-82, 2009.

Zhang J. Evolution by gene duplication: an update. **Trends in Ecology and Evolution**, .v. 18, n.6, 2003.

Zhao L, Pridgeon JW, Becnel JJ, Clark GG, Linthicum KJ. Cytochrome c Gene and Protein Expression: Developmental Regulation, Environmental Response, and Pesticide Sensitivity in *Aedes aegypti*. **Journal of Medical Entomology**, v. 45, p. 401-408, 2008.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo